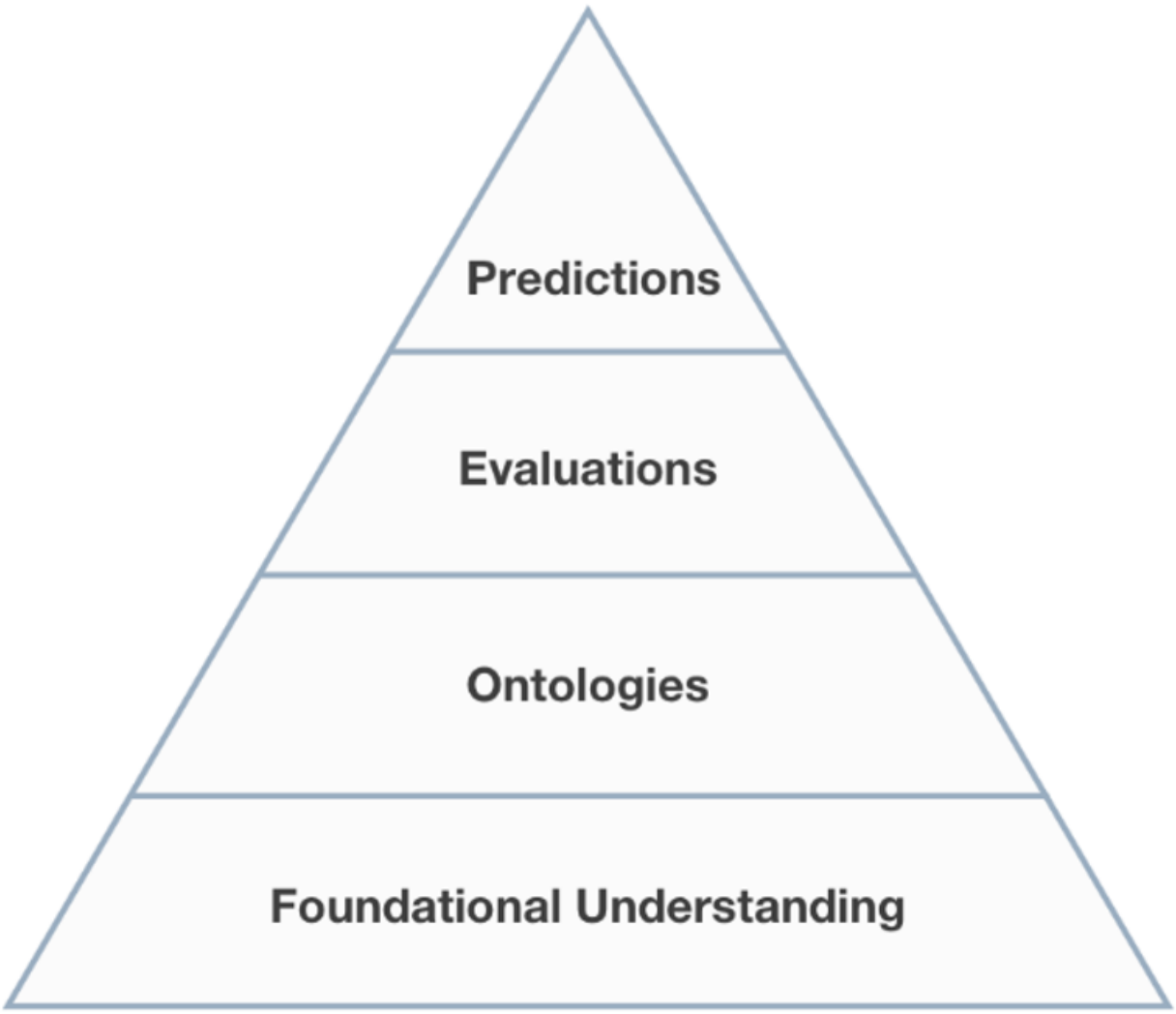
Prediction-Driven Collaborative Reasoning Systems

The Prediction Pyramid: Why Fundamental Work is Needed for Prediction Work

Epistemic state: I feel like this post makes fairly intuitive claims, but I have uncertainty on many of the specifics.

In data science, it is a common mistake for organizations to focus on specific exciting parts like machine learning and data visualizations, while overlooking the infrastructural concerns required to allow for such things. There have been several attempts at making pyramids to showcase the necessary data science dependencies in order to make the most accessible parts realizable.

Similar could be said for predictions. Predictions require foundational work in order to be possible and effective. We can use the prediction pyramid below to show this dependency.



Evaluations

People beginning a prediction practice quickly run into the challenge of having well-specified questions. It's not enough to ask who will win a sports game, one needs to clarify how every exceptional situation is to be handled.[1]

Question specification is a big part of Metaculus. Often questions carry significant discussion even after a question is posed in order to discuss possible edge cases.

In addition to question specification, evaluations can be costly to perform. Even in simple cases it still requires manual work. In more complex cases evaluations could take a very long time. GiveWell does charity evaluations, and these can be extensive. This document discusses some other kinds of evaluations, in the "Possible Uses" section.

Ontologies

Say one is trying to determine which diseases will be important to worry about in 2025. One would first need a taxonomy of diseases that will not change until after 2025. If they were to somehow use a poor or an unusual taxonomy, resulting information wouldn't be useful to others.

In the case of diseases, decades of research years have been carried out in order to establish pragmatic and popular taxonomies. In other domains, new ontologies would need to be developed. Note that we consider ontologies to be a superset of taxonomies.

Another example: the usefulness of careers. 80,000 Hours is an expert here. They have a system which splits career paths into several distinct domains, and rates each one using six distinct attributes. They then do evaluations for each combination.

If it were assured they would continue to do so in the future, it would be relatively straightforward to forecast their future evaluations. If one wanted to do similar predictions without their work, one would have to come up with their own foundational thinking, ontologies, and evaluations.

Other concrete examples of ontologies, for concreteness:

- The "Importance, Neglectedness, Tractability" framework for evaluating charity effectiveness
- Nick Bostrom's typology of information hazards, categorizing them by types and subtypes of information transfer mode and effect
- Nick Bostrom's definition of the vulnerable world hypothesis, including the "semi-anarchic default condition" consisting of limited capacity for preventive policing and global governance, and diverse motivations of actors
- "Posts on LessWrong" are already discrete, and would represent a taxonomy

Foundational Understanding

Even before worrying about predictions or ontologies, it's important to have good foundational understandings of topics in question. An ancient Greek scholar believing in Greek Mythology may spend a lot of time creating ontologies around the gods, but this would be a poor foundation for pragmatic work.

In the case of GiveWell, it took some specific philosophical understanding to decide that charity effectiveness was an important thing to optimize for. Later they came up with the "Importance, Neglectedness, Tractability" framework based on this understanding.

Implications

Predictions are most effective within a cluster of other specific tools.

For predictions to be useful, several other things need to go well, and thus they are also worth paying attention to. Discussions about "doing great predictions" should often include information on these other aspects. The equivalent in data science would be to recognize the importance and challenges of fundamental issues like data warehousing when discussing the eventual goal of data visualization.

Areas with existing substantial fundamental work should be easy to add predictions to.

There are many kinds of data which are already categorized and evaluated; in these cases, the predictions can be quite straightforward. For instance, the "winner of the next presidential election" seems obviously important and will be decided by existing parties, so is a very accessible candidate for forecasting.

It could be good to make lists of metrics and data sources that will be both interesting and reliably provided in the future. For example, it's very likely that Wikidata will continue to report on the GDP and population of different countries, at least for the next 5-10 years. Setting up predictions on such variables should be very feasible.

There could be useful foundational non-predictive work to help future predictions.

One could imagine many useful projects and organizations that focus on just doing a good job on the foundational work, with the goal of assisting predictions down the line. For example, an organization could be set up just to evaluate important future variables. While this organization wouldn't do forecasting itself, it would be very easy for other forecasting efforts to amplify this organization by forecasting its future evaluations. Currently, this is one accidental benefit of some organizations, but if it were intentional then evaluations could be better optimized for prediction support.

Possible Pyramid Modifications

The above pyramid was selected to be a simple demonstration to explain the above implications. In data science, several different pyramids have been made for different circumstances. Similarly, we can imagine multiple variations of this pyramid for other use cases.

"Aggregations" may make sense on top of predictions. It could be possible for some sites to list predictions and others to aggregate them. There are already sites exist to do nothing except for aggregation. Predictwise is one example.

The foundational understanding layer in the bottom could be subdivided into many other categories. For instance, research distillation could be a valid layer.

Acknowledgements

Thanks to Jacob Lagerros for contributing many examples and details to this post, and to Ben Goldhaber and Max Daniel for providing feedback on it.

[1] One of the first markets on prediction market Augur had this exact problem, with no mention of how a sports market would resolve if the game rained out, disputed, postponed, tied, etc. (Zvi discusses this issue further in his post on prediction markets.)

Predictive Reasoning Systems

Meta: This is meant to be a succinct and high-level overview. I expect to go more into detail on parts of the system in future posts. This work builds on this previous post.

Predictions can be a powerful tool but are typically difficult to use in isolation. Being sure to predict the right things is hard. Organizing information from predictions is hard. Coming up with decision options to be predicted is hard. If we seek to create a crowdsourced system of predictions to give us useful information, it would be really useful if we could figure out solutions to these other problems as well.

In the future, I expect that much of the most important prediction work will exist within larger ecosystems of collective reasoning. We can call such systems "Predictive Reasoning Systems."

To give a motivating example, there's been some speculation on Futarchy, a concept of government that involves using prediction markets to determine which government policies would be optimal. Attempts to do this with existing prediction systems could pose substantial challenges. Which specific prediction questions should be asked? How should resources be allocated among questions? The space of all possible policies is gigantic, so how will this best be ideated about and then narrowed down? Predictive Reasoning Systems outline a high-level construct that may be more capable of handling such high-level challenges.

Prediction Systems vs. Predictive Reasoning Systems

We can call a system used exclusively to make predictions a Prediction System.

One interacts with a Prediction System by posing and answering prediction questions. Points or subsidization can be assigned by users to prioritize work on what seems like the most useful questions.

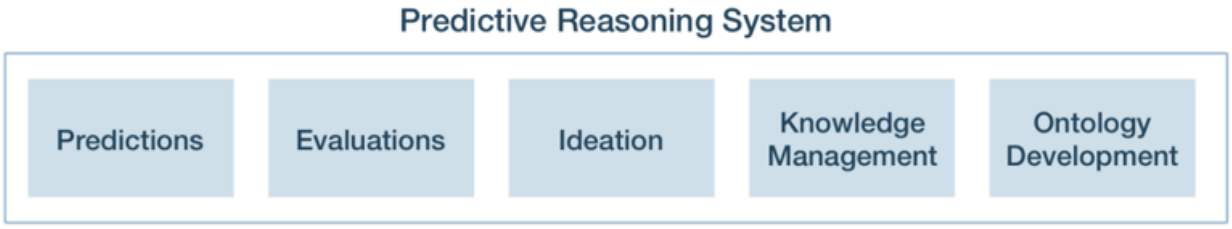
In contrast, one would interact with a Predictive Reasoning System by giving it higher level goals. For example, "Optimally spend 1000 human hours to come up with whatever information would most help our business." The system would translate these goals into lots of lower level work, including the use of forecasting.

If a group of people poses multiple questions on a Prediction System, with one overarching goal in mind, one can refer to this [combination of Prediction System + question askers] as a simple Predictive Reasoning System. One can even say that wherever a Prediction System exists, a Predictive Reasoning System of some sort exists around it. That said, just because it formally exists does not mean that it's well optimized.

Primary Functions

There many possible ways to categorize the functionality and subcomponents of Predictive Reasoning Systems, but a good one to start with may be to categorize them by high-level function. One attempt at this would split functions up as predictions,

evaluations, ideation, knowledge management, and ontology development. Each of these encompasses a great deal of existing thought and literature, but could still use considerable additional work for best use in Predictive Reasoning Systems.



Predictions

Quite obviously, predictive reasoning systems should make predictions. There should be many predictable measures for forecasters to work on.

Evaluations

Many predictable questions should eventually be evaluated (judged) in order to measure predictor performance and possibly reward them accordingly. There are many different ways of doing this.

Human-involved evaluations present a challenging problem with substantial literature outside of forecasting. Trustworthy oracles of even simple parameters also present challenges that have discussed most recently around blockchain applications.

Example Evaluations:

"What will the GDP be of the United States in 2030, according to Wikipedia?"

"What will the next Quinnipiac poll rate the United States president in 2024?"

"What will be the estimated counterfactual monetary value of project X, according to independent auditor Y?"

Ideation

In some cases decisions and things to predict are obvious. In many others, especially as things scale, they aren't.

For example, say a business is considering a specific project proposal. They can either rejected or accept it. In this case, there is no immediate ideation necessary. Predictors could predict the outcome (such as profit in the next quarter) conditional on the business rejecting or accepting the proposal.

But this is not a very realistic scenario. In many cases the proposal could be modified, in which case, there could be a very large space of accessible options. Here it could be very nice to have some very knowledgeable and creative people suggesting improvements. If this were combined with a flexible prediction system it may be able to converge upon a much better proposal than either predictions or ideation could independently end up with.

Crowd ideation has been one of the most successful use cases of online crowdsourcing, so could be a natural component of an online crowdsourced prediction system.

Knowledge Management

Competent forecasters learn reusable facts and practices during their work. They may also be able to be augmented with non-forecast researchers to help collect and provide useful information. Financial traders typically work with an extensive support staff of operations and research assistance, and it would make sense that great forecasters would be aided by similar help.

On a related note, knowledge management systems are highly valuable to most organizations, so it would make sense that they would also be useful for groups of forecasters working on similar subjects.

With some structure, knowledge management should not only be useful to forecasting, but forecasting could also be useful to knowledge management. Forecasts could be made on the benefits of specific knowledge modifications, leading to increasingly efficient improvements as a Predictive Reasoning System becomes more advanced.

Ontology Development

In order to forecast the most useful things, it's important to organize information and forecasts into reasonably effective ontologies. This includes theoretical domain work in order to arrive at and properly understand the most effective ontologies.

If an organization would want to identify the most useful business practices to adopt, it would first have to establish a decent taxonomy of the possible options. If it were to do a poor job with this and then spent a lot of effort forecasting, and then later dramatically change its taxonomy, then much of that work could be wasted.

In another frame, ontology selection is very similar to the issue of feature selection in data science. A poor choice of variables to predict would lead to results that both aren't decision relevant nor otherwise interesting.

Optional Functions

The above functions are those that seemed most necessary for Predictive Reasoning Systems in the next 5-15 years. However, there's one more that wouldn't change the reasoning structure but would modify behavior.

Action Follow-Through (Input & Output)

Predictive Reasoning Systems as stated above are meant to create as much value as possible in the form of information, specifically by using predictions. The idea is that the result of this work will be used for future decisions and that those decisions will be made by agents external to these systems. However, that doesn't exactly have to be the case. With some minor changes, it would be possible for Predictive Reasoning Systems to invoke direct agency and trigger actions in the world.

For example, one system may determine, with high-certainty, that one good course of action for an organization would be to purchase a new lamp. This could automatically

trigger an online order.

The capability of Predictive Reasoning Systems to trigger direct actions in the world is here called "Action Follow-Through", and can be thought of as similar to I/O in software systems. It can raise significant safety issues, though it will probably be a while before it ever becomes practical. That said, if it becomes practical, it is possible that Predictive Reasoning Systems with this capability may be much more powerful than ones without it.

Predictive Reasoning Systems with Action Follow-Through could be considered one form of "decentralized autonomous corporation".

Future Work

This document lays out a definition of Predictive Reasoning Systems with very simple descriptions of what currently seems like their main functions. Each of these functions has a rich literature and considerable room for more thought. There could be a good deal of work categorizing this literature and researching the most promising methods in each function for the use in these predictive systems. There are also other categorization systems to better explore the solution space of Predictive Reasoning Systems. I expect to address some of these topics in future work.

Thanks to Ben Goldhaber and Jacob Lagerros for feedback on this post.

Prediction-Augmented Evaluation Systems

[Note: I made a short video of myself explaining this document here.]

It's common for groups of people to want to evaluate specific things. Here are a few examples I'm interested in:

- The expected value of projects or actions within projects
- Research papers, on specific rubrics
- Quantitative risk estimates
- Important actions that may get carried out by artificial intelligences

I think predictions could be useful in scaling and amplifying such evaluation processes. Humans and later AIs could predict intensive evaluation results. There has been previous discussion on related topics, but I thought it would be valuable to consider a specific model here called "prediction-augmented evaluation processes." This is a high-level concept that could be used to help frame future discussion.

Desiderata:

We can call a systematized process that produces evaluations an "evaluation process." Let's begin with a few generic desiderata of these.

- **High Accuracy / "Evaluating the right thing"**
  - Evaluations should aim at estimating the thing actually cared about as well as possible. In their limit according to some metric of effort, they should approximate ideal knowledge on the thing cared about.
- **High Precision / "Evaluating the chosen thing correctly"**
  - Evaluations should have low amounts of uncertainty and be very consistent. If the precision is generally less than what naive readers would guess, then these evaluations wouldn't be very useful.
- **Low Total Cost**
  - Specific evaluations can be costly, but the total cost across evaluations should be low.

I think that the use of predictions could allow us to well fulfill these criterions. It could help decouple evaluations from their scaling, allowing for independent optimization of the first two. The cost should be low relative to that of scaling evaluators in other obvious ways.

Prediction-Augmentation Example

Before getting formal with terminology, I think a specific example would be helpful.

Say Samantha scores research papers for quality on a scale from 1-10. She's great at it, she has a very thorough and lengthy reviewing procedure, and many others trust her reviews. Unfortunately, there's only one Samantha, and there are tons of research papers.

One way to scale Samantha's abilities would be to use a prediction aggregation system. A collection of other people would predict Samantha's scores before she rates them. Predictions would be submitted as probability distributions over possible scores. Each research paper would have a probability of being scored by Samantha, say 10%. In a naive model, this would be done in batches; the predictors could have 1 month to score 100 papers, and then at the end of the month 10 would randomly be chosen and rated by Samantha.

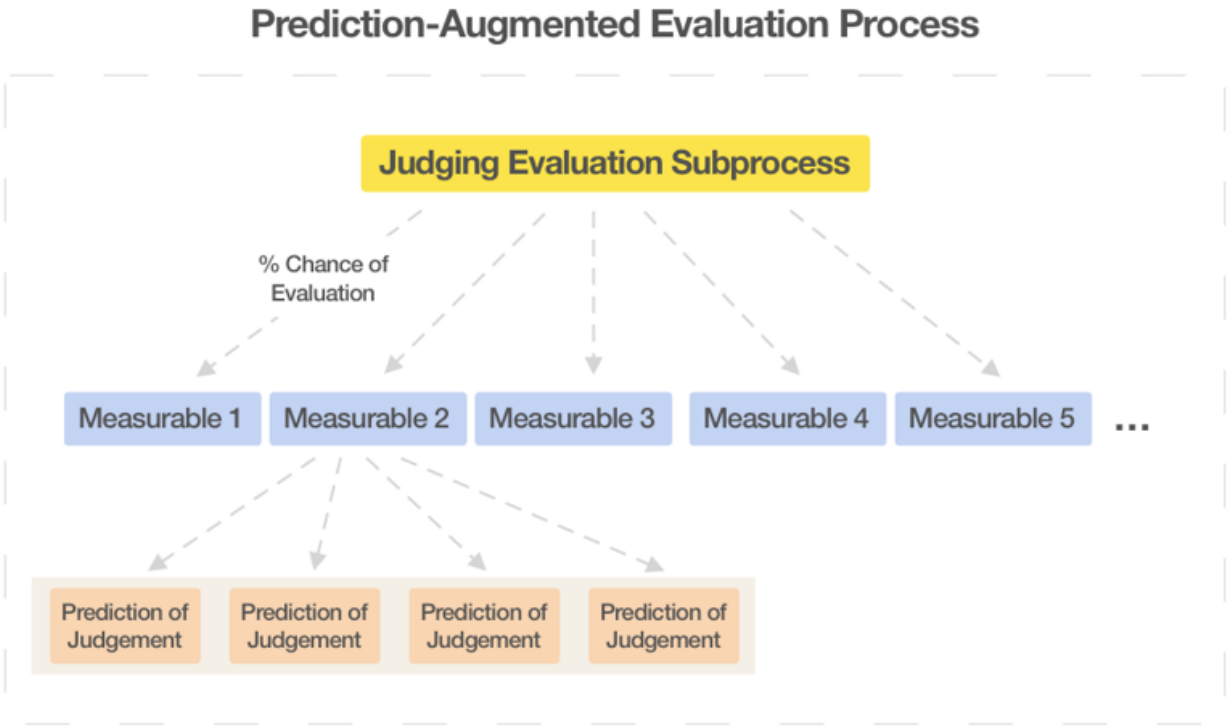
If this batch process would happen multiple times, then eventually outside observers could understand how accurate the predictors are and how to aggregate future forecasts to better predict Samantha's judgments.

An obvious improvement could be that some of the predictors may develop a sense of what arguments Samantha most likes and what data she cares for. They may write up summaries of their arguments to convince Samantha of their particular stances. If managed well, this could speed up Samantha's work and perhaps improve it. She may eventually find many of the people who best understand her system and develop an amount of trust in them. Of course, this could selectively bias her away from making accurate judgments, so this kind of feedback would have to be handled with care.

Once there are enough predictions, it may be possible to train ML agents to do prediction as well. The humans would essentially act as a "bootstrapping" system.

Subcomponents

I've outlined how I would describe the internals of a prediction-augmented evaluation process in an engineering system or similar. The wording here is a bit technical, on purpose, so feel free to skip this section.



This diagram attempts to show a few different things. The entirety of a judging evaluation subprocess and prediction system make up the outer prediction-augmented evaluation process. The judging evaluation subprocess has a percent chance of evaluating each of a set of measurables. Predictors can make predictions on each one of these measurables, where they are trying to predict what the judging evaluation subprocess will judge for that measurable if it's chosen to judge it.

Judging Evaluation Subprocess

I imagine that prediction-augmentation could assist any evaluation process, even theoretically one that is already itself prediction-augmented. Prediction-augmentation acts as a layer that converts one narrow but good evaluation process into a more voluminous process.

In the context of a "prediction-augmented" evaluation process, the "wrapped" evaluation process can be considered the "judging" evaluation subprocess. This internal process would generate "judgments", and separately predictors will make predictions of future judgments. Both judgments and predictions would act as evaluations, so to speak.

There are already many evaluation systems used in the world, and I imagine that almost any could act as judging processes. The main bottlenecks would be judging quantity and reliability; this would be most useful for areas where evaluations are done for many similar things.

Because the judging process is well isolated, and scale is not a huge worry (that's pushed to the prediction layer), it can be thoroughly tested and optimized. Because the scaling mechanism is decently decoupled from the evaluation process, it could be much more rigorous than would otherwise be reasonable. For instance, a paper reviewer may typically spend 4 hours per paper, but with a prediction-augmented layer, perhaps they could spend 40 with the papers selected for judgment.

I use the phrase "evaluation process" rather than "evaluation" to point out the fact that this should be something outside the purview of a single individual. I imagine that the failure rate of individuals to evaluate things after a few years could be considerable, so it would be strongly preferable to have backup plans in the case that that happens. I would assume that organizations would generally be a better alternative, even if they were just mostly backing up individuals. Perhaps organizations could set up official trusts or other legal and financial structures to ensure that judgments get carried out.

There would have to be discussion about what the best evaluation processes would look like if many resources were put into predictions, but I think that's a really good discussion to encourage anyway.

One tricky part would be to further identify evaluation processes that multiple agents would find most informative. For instance, finding some individual that's trusted by several organizations with significant differences of opinion.

Measurables

Measurables refer to the things that get evaluated. It's a bit of a generic word for the use case, but I suspect useful in larger ontologies. Some examples could be "the rating of scientific paper X" or "the expected value of project Y." It's important to keep in mind that measurables only make sense in regards to specific evaluation systems; predictors would rarely predict the actual value of something, but rather, the result of a specific

evaluation subprocess. For instance, "GDP of the United States, according to XYZ's process."

Predictions
The system obviously requires predictions, and for this to happen at a decent scale, almost definitely some kind of web application. In theory, a formal prediction market would work, but I imagine it would be very difficult to scale to the levels I would hope for in a large evaluation system. I'm personally more excited about more general prediction aggregation tools like The Good Judgment Project and Metaculus. Metaculus, in particular, allows participants to make guesses on continuous variables, which seems like a reasonable mechanism for evaluation systems. I'm also experimenting with a small project of my own to collect forecasts for experimental purposes.

Incentives for predictors could be a bit tricky to work out, but it definitely seems possible. It seems simple enough to pay people using a function that includes their prediction accuracy and quantity. Sign-ups could be screened to prevent lots of bots from joining. Of course, another option would be for the benefits from predictors to be something that itself gets evaluated using a separate prediction-augmented process.

Scaling & Amplification

I think the main two benefits Prediction-Augmentation could provide are that of "scaling" and "amplification." "Scaling" refers to the ability of such a system to effectively "scale" an evaluation judgment subprocess. The predictors would evaluate many more measurables than the judgment subprocess, and would do so sooner. "Amplification" refers to the ability of the system to improve the best abilities of the judging subprocess. This could come from speeding it up and/or by having judges read content produced by the prediction layer.

I expect "scaling" to be much more impactful than "amplification," especially for the early use of such systems.

Scaling & Amplification are very similar in ways to "Iterated Distillation and Amplification." However, these types of scaling & amplification are obviously not always automated, which is a big difference. That said, hypothetically people could eventually write prediction bots, and similar ones for amplification (with nice user interfaces, I assume.) I think prediction-augmentation may have relevance for direct use in technical AI alignment systems but I am currently more focused on human variants.

Existing/Possible Variants

Selective Evaluations
The judgment subprocess could select specific predicted variables for evaluations after reviewing the predictions, rather than choosing probabilistically. Judges would essentially "challenge" the measurables with the most questionable predictions. Selective evaluations may be more efficient than random evaluations, though it also could mean that predictors may be incentivized to predict items they expect the evaluators would select, leading to some potentially messy issues.

Selective evaluation is essentially very similar to some things many editors and managers do. A news editor may skim a long work by a writer (who is acting in part as

a predictor of what the editor will accept), and at times challenge specific parts of text, to either improve directly or send back for improvement.

EV-Adjusted Probabilities
If evaluations are done probabilistically, the probabilities could change depending on the expected value of improved predictions on specific measurables. This could incentivize the predictors to allocate more effort accordingly. This could look a lot like selective evaluations in practice.

Traditional Prediction Systems
I would consider existing prediction aggregators/markets to fall under the umbrella of "Prediction-Augmented Evaluation Processes." These traditionally have had judging subprocesses that are very straightforward and simple; for instance, "Find the GDP of America in 2020 from Wikipedia." They effectively scale simple judgments purely by estimating them early, rather than also by attempting to recreate a complicated analysis.

Possible Uses

Project Evaluations
Projects could be evaluated for their expected marginal impact. This could provide information very similar to certificates of impact. I think that prediction-augmented evaluation systems could be more efficient than certificates of impact, but would first like to see both be tested more experimentally. This post by Ought poses a similar system for doing evaluations on parts of projects. This post by Robin Hanson discusses similar techniques for evaluating the impact of scientific papers.

General Research Questions
If researchers could express specific uncertain claims early on, then outsiders could predict these researcher's eventual findings. For example, a scientist could make a list of 100 binary questions they are not sure about, and promise to evaluate a random subset in 10 years.

AI Decision Validation
One possibility here could be to have a human act as a judge (hopefully augmented in some way), and an intelligent AI be the predictor. The AI would recommend actions/decisions to the human, and the human/augmentation system would selectively or statistically challenge these. I believe this is similar to ideas of selective challenging in AI Safety via Debate.

Human Value Judgement
If we could narrow value judgments into a robust evaluation process, we could scale this to AI systems. This could be used for making decisions around self-driving vehicles and similar. I imagine that much of the challenge here would be for people to agree on evaluation processes for moral questions, but if this could be approximated, the rest could be carried out somewhat straightforwardly. See this post by Paul Christiano for more information.

Website Moderation
Many forums and applications are pretty dependent on specific moderators for moderation. This kind of work could hypothetically help scale them in a controllable way. Future moderators would be obligated to predict the trusted moderators, rather than doing things in other ways. I'm not too sure about this, but know that others in the

community have been enthusiastic. See this post by Paul Christiano for more information.

Alternative Dispute Resolution
Existing court systems and alternative dispute resolution systems already are similar to this process in theory. It would be interesting to imagine hypothetical court systems where lower courts would try to predict exactly what higher courts would rule, and on occasion, the higher courts would repeat the same cases. The appellate system may be more efficient, but there may be interesting hybrids. For one, this system could be useful for bootstrapping completely automated rulings.

Unimagined Uses
I imagine many of the most interesting uses of such a system haven't thought about. Prediction-augmented evaluation processes would have some positives and negatives current systems don't have, so may make sense in different cases. If they do very well, I would assume they may do so in ways that would surprise us.

Related Work

Much of what has been discussed here is very generic and thus many parts have been previously considered. Paul Christiano, and the team of Ought, in particular, have written about very similar ideas before; the main difference is that they seem to have focussed more on AI learning and specific decisions. Ought's Predicting Slow Judgements" work investigates how well humans make predictions on different scales of time for evaluations, and then how that could be mimicked by AIs. I've done some work with them before and recommend them to others interested in these topics. Andreas Stuhlmüller's (founder of Ought) previous work with dialog markets is also worth reading.

There seems to be a good amount of research on evaluation procedures and separately on prediction capabilities. For the sake of expediency, I did not treat this as much of a literature review, though would be interested in whether others have recommended literature on these topics.

What if people simply forecasted your future choices?

tldr: If you could have a team of smart forecasters predicting your future decisions & actions, they would likely improve them in accordance with your epistemology. This is a very broad method that's less ideal than more reductionist approaches for specific things, but possibly simpler to implement and likelier to be accepted by decision makers with complex motivations.

Background

The standard way of finding questions to forecast involves a lot of work. As Zvi noted, questions should be very well-defined, and coming up with interesting yet specific questions takes considerable consideration.

One overarching question is how predictions can be used to drive decision making. One recommendation (one version called "Decision Markets") often comes down to estimating future parameters, conditional on each of a set of choices. Another option is to have expert evaluators probabilistically evaluate each option, and have predictors predict their evaluations (Prediction-Augmented Evaluations.)

Proposal

One prediction proposal I suggest is to **have predictors simply predict the future actions & decisions of agents**. I temporarily call this an **"action prediction system."** The evaluation process (the choosing process) would need to happen anyway, and the question becomes very simple. This may seem too basic to be useful, but I think it may be a lot better than at least I initially expected.

Say I'm trying to decide what laptop I should purchase. I could have some predictors predicting which one I'll decide on. In the beginning, the prediction aggregation shows that I have an 90% chance of choosing one option. While I really would like to be the kind of person who purchases a Lenovo with Linux, I'll probably wind up buying another Macbook. The predictors may realize that I typically check Amazon reviews and the Wirecutter for research, and they have a decent idea of what I'll find when I eventually do.

It's not clear to me how to best focus predictors on specific uncertain actions I may take. It seems like I would want to ask them mostly about specific decisions I am uncertain of.

One important aspect is that I should have a line of communication to the predictors. This means that some clever ones may eventually catch on to practices such as the following:

A forecaster-sales strategy

1. Find good decision options that have been overlooked

2. Make forecasts or bets on them succeeding

3. Provide really good arguments and research as to why they are overlooked

If I, the laptop purchaser, am skeptical, I could ignore the prediction feedback. But if I repeat the process for other decisions eventually I should eventually develop a sense of trust in the aggregation accuracy, and then in the predictor ability to understand my desires. I may also be very interested in what that community has to say, as they have developed a model of what my preferences are. If I'm generally a reasonable and intelligent person, I could learn how to best rely on these predictors to speed up and improve my future decisions.

In a way, this solution doesn't solve the problem of "how to decide the best option;" it just moves it into what may be a more manageable place. Over time I imagine that new strategies may emerge for what generally constitutes "good arguments", and those will be adopted. In the meantime, agents will be encouraged to quickly choose options they would generally want, using reasoning techniques they generally prefer. If one agent were really convinced by a decision market, then perhaps some forecasters would set one up in order to prove their point.

Failure Modes

There are few obvious failure modes to such a setup. I think that it could dilute signal quality, but am not as worried about some of the other obvious ones.

Weak Signals

I think it's fair to say that if one wanted to optimize for expected value, asking forecasters to predict actions instead could lead to weaker signals. Forecasters would be estimating a few things at once (how good an option is, and how likely the agent is to choose it.) If the agent isn't really intent on optimizing for specific things, and even if they are, it may be difficult to provide enough signal in their probabilities of chosen decisions for them to be useful. I think this would have to be empirically tested under different conditions.

There could also be complex feedback loops, especially for naive agents. An agent may trust its predictors too much. If the predictors believe the agent is too trusting or trusts the wrong signals, they could amplify those signals and find "easy stable points." I'm really unsure of how this would look or how much competence the agent or predictors would need to have net-beneficial outcomes. I'd be interested in testing and paying attention to this failure mode.

That said, the reference class of groups who were considering and interested in paying for using "action predictions" vs. "decision markets" or similar is a very small one, and one that I expect would be convinced only by pretty good arguments. So pragmatically, in the rare cases where the question of "would our organization be wise enough to get benefit from action predictions" is asked, I'd expect the answer to lean positively. I wouldn't expect obviously sleazy sales strategies to work to convince GiveWell of a new top cause area, for example.

Inevitable Failures

Say the predictors realized that a MacBook wouldn't make any sense for me, but that I was still 90% likely to choose it, even after I heard all of the best arguments. It would be somewhat of an "inevitable failure." The amount of utility I get from each item could be very uncorrelated with my chances of choosing that item, even after hearing about that difference.

While this may be unfortunate, it's not obvious what *would* work in these conditions. The goal of predictions shouldn't be to predict the future accurately, but instead to help agents make better decisions. If there were a different system that did a great job outlining the negative effect of a bad decision to my life, but I predictably ignored the system, then it just wouldn't be useful, despite being accurate. Value of information would be low. It's really tough for a system of information to be so good as to be useful even when ignored.

I'd also argue that the kinds of agents that would make predictably poor decisions would be ones that really aren't interested in getting accurate and honest information. It could seem pretty brutal to them; basically, it would involve them paying for a system that continuously tells them that they are making mistakes.

This previous discussion has assumed that the agents making the decisions are the same ones paying for the forecasting. This is not always the case, but in the counterexamples, setting up other proposals could easily be seen as hostile. If I set up a system to start evaluating the expected total values of all the actions of my friend George, knowing that George would systematically ignore the main ones, I could imagine George may not be very happy with his subsidized evaluations.

Principal-agent Problems

I think "*action predictions*" would help agents fulfill their actual goals, while other forecasting systems would more help them fulfill their stated goals. This has obvious costs and benefits.

Let's consider a situation with a CEO who wants to their company to be as big as possible, and corporate stakeholders who want instead for the company to be as profitable as possible.

Say the CEO commits to "maximizing shareholder revenue," and commits to making decisions that do so. If there were a decision market set up to tell how much "shareholder value" would be maximized for each of a set of options (different to a decision prediction system), and that information was public to shareholders, then it would be obvious to them when and how often the CEO disobeys that advice. This would be a very transparent set up that would allow the shareholders to police the CEO. It would take away a lot of flexibility and authority of the CEO and place it in the hands of the decision system.

On the contrary, say the CEO instead shares a transparent action prediction system. Predictor participants would, in this case, try to understand the specific motivations of the CEO and optimize their arguments as such. Even if they were being policed by shareholders, they could know this, and disguise their arguments accordingly. If discussing and correctly predicting the net impact to shareholders would be net harmful in terms of predicting the CEO's actions and convincing them as such, they could simply ignore it, or better yet find convincing arguments not to take that action. I expect that an action prediction system would essentially act to amplify the abilities of the decider, even if at the cost of other caring third parties.

Salesperson Melees

One argument against this is a gut reaction that it sounds very "salesy", so probably won't work. While I agree there are some cases where it may not too work well (stated above in the weak signal section), I think that smart people should be positively augmented by good salesmanship under reasonable incentives.

In many circumstances, salespeople practically are really useful. The industry is huge, and I'm under the impression that at least a significant fraction (>10%) is net-beneficial. Specific kinds of technical and corporate sales come to mind, where the "sales" professionals are some of the most useful for discussing technical questions with. There simply aren't other services willing to have lengthy discussions about some topics.

Externalities

Predictions used in this way would help the goals of the agents using them, but these agents may be self-interested, leading to additional negative externalities on others. I think this prediction process doesn't at all help in making people more altruistic. It simply would help agents better satisfy their own preferences. This is a common aspect to almost all intelligence-amplification proposals. I think it's important to consider, but I'm really recommending this proposal more as a "possible powerful tool", and not as a "tool that is expected to be highly globally beneficial if used." That would be a very separate discussion.

Can We Place Trust in Post-AGI Forecasting Evaluations?

TLDR
Think "A prediction market, where most questions are evaluated shortly after an AGI is developed." We could probably answer hard questions more easily post-AGI, so delaying them would have significant benefits.

Motivation

Imagine that select pre-AGI legal contracts stay valid post-AGI. Then a lot of things are possible.

There are definitely a few different scenarios out there for economic and political consistency post-AGI, but I believe there is at least a legitimate chance (>20%) that legal contracts will exist for what seems like a significant time (>2 human-experiential years.)

If these contracts stay valid, then we could have contracts set up to ensure that prediction evaluations and prizes happen.

This could be quite interesting because post-AGI evaluations could be a whole lot better than pre-AGI evaluations. They should be less expensive and possibly far more accurate.

One of the primary expenses now with forecasting setups is the evaluation specification and execution. If these could be pushed off while keeping relevance, that could be really useful.

Idea

What this could look like is something like a Prediction Tournament or Prediction Market where many of the questions will be evaluated post-AGI. Perhaps there would be a condition that the questions would only be evaluated if AGI happens within 30 years, and in those cases, the evaluations would happen once a specific threshold is met.

If we expect a post-AGI world to allow for incredible reasoning and simulation abilities, we could assume that it could make incredibly impressive evaluations.

Some example questions:

- To what degree is each currently-known philosophical system accurate?
- What was the expected value of Effective Altruist activity Y, based on the information available at the time to a specific set of humans?
- How much value has each Academic field created, according to a specific philosophical system?
- What would the GDP of the U.S. have been in 2030, conditional on them doing policy X in 2022?

- What were the chances of AGI going well, based on the information available at the time to a specific set of humans?

Downsides

My guess is that many people would find this quite counterintuitive. Forecasting systems are already weird enough.

There's a lot of uncertainty around the value systems and epistemic l states of authoritative agencies, post-AGI. Perhaps they would be so incredibly different to us now that any answers they could give us would seem arcane and useless. Similar to how it may become dangerous to extrapolate one's volition "too far", it may also be dangerous to be "too smart" when making evaluations defined by less intelligent beings.

That said, the really important thing isn't how the evaluations will actually happen, but rather what forecasters will think of it. Whatever evaluation system motivates forecasters to be as accurate and useful as possible (while minimizing cost) is the one to strive for.

My guess is that it's worth trying out, at least in a minor capacity. There should, of course, be related forecasts for things like, "In 2025, will it be obvious that post-AGI forecasts are a terrible idea?"

Questions for Others

This all leaves a lot of questions open. Here are a few specific ones that come to mind:

- What kinds of legal structures could be most useful for post-AGI evaluations?
- What, in general, would people think of post-AGI evaluations? Could any prediction community take them seriously and use them for additional accuracy?
- What kinds of questions would people want to see forecasted, if we could have post-AGI evaluations?
- What other factors would make this a good or bad thing to try out?

Ideas for Next Generation Prediction Technologies

2021 Note: This was written and posted in December, 2016. The date it shows on LessWrong is 2019; I believe this refers to a time the post was (very minorly) updated, as part of moving it to the Prediction-Driven Collaborative Reasoning Systems sequence.

Prediction markets are powerful, but also still quite niche. I believe that part of this lack of popularity could be solved with significantly better tools. During my work with Guesstimate I've thought a lot about this issue and have some ideas for what I would like to see in future attempts at prediction technologies.

1. Machine learning for forecast aggregation

In financial prediction markets, the aggregation method is the market price. In non-market prediction systems, simple algorithms are often used. For instance, in the Good Judgement Project, the consensus trends displays "the median of the most recent 40% of the current forecasts from each forecaster."[1] Non-financial prediction aggregation is a pretty contested field with several proposed methods.[2][3][4]

I haven't heard much about machine learning used for forecast aggregation. It would seem to me like many, many factors could be useful in aggregating forecasts. For instance, some elements of one's social media profile may be indicative of their forecasting ability. Perhaps information about the educational differences between multiple individuals could provide insight on how correlated their knowledge is.

Perhaps aggregation methods, especially with training data, could partially detect and offset predictable human biases. If it is well known that people making estimates of project timelines are overconfident, then this could be taken into account. For instance, someone enters in "*I think I will finish this project in 8 weeks*", and the system can infer something like, "*Well, given the reference class I have of similar people making similar calls, I'd expect it to take 12.*"

A strong machine learning system would of course require a lot of sample data, but small strides may be possible with even limited data. I imagine that if data is needed, lots of people on platforms like Mechanical Turk could be sampled.

2. Prediction interval input

The prediction tools I am familiar with focus on estimating the probabilities of binary events. This can be extremely limiting. For instance, instead of allowing users to estimate what Trump's favorable rating would be, they instead have to bet on whether it will be over a specific amount, like "Will Trump's favorable rate be at least 45.0% on December 31st?"[5]

It's probably no secret that I have a love for probability densities. I propose that users should be able to enter probability densities directly. User entered probability densities would require more advanced aggregation techniques, but is doable.[6]

Probability density inputs would also require additional understanding from users. While this could definitely be a challenge, many prediction markets already are quite complicated, and existing users of these tools are quite sophisticated.

I would suspect that using probability densities could simplify questions about continuous variables and also give much more useful information on their predictions. If there are tail risks these would be obvious; and perhaps more interestingly, probability intervals from prediction tools could be directly used in further calculations. For instance, if there were separate predictions about the population of the US and the average income, these could be multiplied to have an estimate of the total GDP (correlations complicate this, but for some problems may not be much of an issue, and in others perhaps they could be estimated as well).

Probability densities make less sense for questions with a discrete set of options, like predicting who will win an election. There are a few ways of dealing with these. One is to simply leave these questions to other platforms, or to resort back to the common technique of users estimating specific percentage likelihoods in these cases. Another is to modify some of these to be continuous variables that determine discrete outcomes; like the number of electoral college votes a U.S. presidential candidate will receive. Another option is to estimate the ‘true’ probability of something as a distribution, where the ‘true’ probability is defined very specifically. For instance, a group could make probability density forecasts for the probability that the blog 538 will give to a specific outcome on a specific date. In the beginning of an election, people would guess 538's percent probability for one candidate winning a month before the election.

3. Intelligent Prize Systems

I think the main reason why so many academics and rationalists are excited about prediction markets is because of their positive externalities. Prediction markets like InTrade seem to do quite well at predicting many political and future outcomes, and this information is very valuable to outside third parties.

I’m not sure how comfortable I feel about the incentives here. The fact that the main benefits come from externalities indicates that the main players in the markets aren’t exactly optimizing for these benefits. While users are incentivized to be correct and calibrated, they are not typically incentivized to predict things that happen to be useful for observing third parties.

I would imagine that the externalities created by prediction tools would be strongly correlate with the value of information to these third parties, which does rely on actionable and uncertain decisions. So if the value of information from prediction markets were to be optimized, it would make sense that these third parties have some way of ranking what gets attention based on what their decisions are.

For instance, a whole lot of prediction markets and related tools focus heavily on sports forecasts. I highly doubt that this is why most prediction market enthusiasts get excited about these markets.

In many ways, promoting prediction markets for their positive externalities is very strange endeavor. It’s encouraging the creation of a marketplace because of the expected creation of some extra benefit that no one directly involved in that marketplace really cares about. Perhaps instead there should be otherwise-similar

ways for those who desire information from prediction groups to directly pay for that information.

One possibility that has been discussed is for prediction markets to be subsidized in specific ways. This obviously would have to be done carefully in order to not distort incentives. I don't recall seeing this implemented successfully yet, just hearing it be proposed.

For prediction tools that aren't markets, prizes can be given out by sponsoring parties. A naive system is for one large sponsor to sponsor a 'category', then the best few people in that category get the prizes. I believe something like this is done by Hypermind.

I imagine a much more sophisticated system could pay people as they make predictions. One could imagine a system that numerically estimates how much information was added to the new aggregate when a new prediction is made. Users with established backgrounds will influence the aggregate forecast significantly more than newer ones, and thus will be rewarded proportionally. A more advanced system would also take into account estimate supply and demand; if there are some conditions where users particularly enjoy adding forecasts, they may not need to be compensated as much for these, despite the amount or value of information contributed.

On the prize side, a sophisticated system could allow various participants to pool money for different important questions and time periods. For instance, several parties put down a total of $10k on the question 'what will the US GDP be in 2020', to be rewarded over the period of 2016 to 2017. Participants who put money down could be rewarded by accessing that information earlier than others or having improved API access.

Using the system mentioned above, an actor could hypothetically build up a good reputation, and then use it to make a biased prediction in the expectation that it would influence third parties. While this would be very possible, I would expect it to require the user to generate more value than their eventual biased prediction would cost. So while some metrics may become somewhat biased, in order for this to happen many others would become improved. If this were still a problem, perhaps forecasts could make bets in order to demonstrate confidence (even if the bet were made in a separate application).

4. Non-falsifiable questions

Prediction tools are really a subset of estimation tools, where the requirement is that they estimate things that are eventually falsifiable. This is obviously a very important restriction, especially when bets are made. However, it's not an essential restriction, and hypothetically prediction technologies could be used for much more general estimates.

To begin, we could imagine how very long term ideas could be forecasted. A simple model would be to have one set of forecasts for what the GDP will be in 2020, and another for what the systems' aggregate will think the GDP is in 2020, at the time of 2018. Then in 2018 everyone could be ranked, even though the actual event has not yet occurred.

In order for the result in 2018 to be predictive, it would obviously require that participants would expect future forecasts to be predictive. If participants thought everyone else would be extremely optimistic, they would be encouraged to make optimistic predictions as well. This leads to a feedback loop that the more accurate the system is thought to be the more accurate it will be (approaching the accuracy of an immediately falsifiable prediction). If there is sufficient trust in a community and aggregation system, I imagine this system could work decently, but if there isn't, then it won't.

In practice, I would imagine that forecasters would be continually judged as future forecasts are contributed that agree or disagree with them, rather than only when definitive events happen that prove or disprove their forecasts. This means that forecasters could forecast things that happen in very long time horizons, and still be ranked based on their ability in the short term.

Going more abstract, there could be more abstract poll-like questions like, "How many soldiers died in war in WW2?" or "How many DALYs would donating $10,000 to the AMF create in 2017?". For these, individuals could propose their estimates, then the aggregation system would work roughly like normal to combine these estimates. Even though these questions may never be known definitively, if there is built in trust in the system, I could imagine that they could produce reasonable results.

One question here which is how to evaluate the results of aggregation systems for non-falsifiable questions. I don't imagine any direct way but could imagine ways of approximating it by asking experts how reasonable the results seem to them. While methods to aggregate results for non-falsifiable questions are themselves non-falsifiable, the alternatives also are very lacking. Given how many of these questions exist, it seems to me like perhaps they should be dealt with; and perhaps they can use the results from communities and statistical infrastructure optimized in situations that do have answers.

Conclusion

Each one of the above features could be described in much more detail, but I think the basic ideas are quite simple. I'm very enthusiastic about these, and would be interested in talking with anyone interested in collaborating on or just talking about similar tools. I've been considering attempting a system myself, but first want to get more feedback.

1. The Good Judgement Project FAQ, https://www.gjopen.com/faq
2. Sharpening Your Forecasting Skills, Link
3. IARPA Aggregative Contingent Estimation (ACE) research program https://www.iarpa.gov/index.php/research-programs/ace
4. The Good Judgement Project: A Large Scale Test of Different Methods of Combining Expert Predictions Link
5. "Will Trump's favorable rate be at least 45.0% on December 31st?" on PredictIt (Link).

6. I believe Quantile Regression Averaging is one way of aggregating prediction intervals https://en.wikipedia.org/wiki/Quantile_regression_averaging

7. Hypermind (http://hypermind.com/)