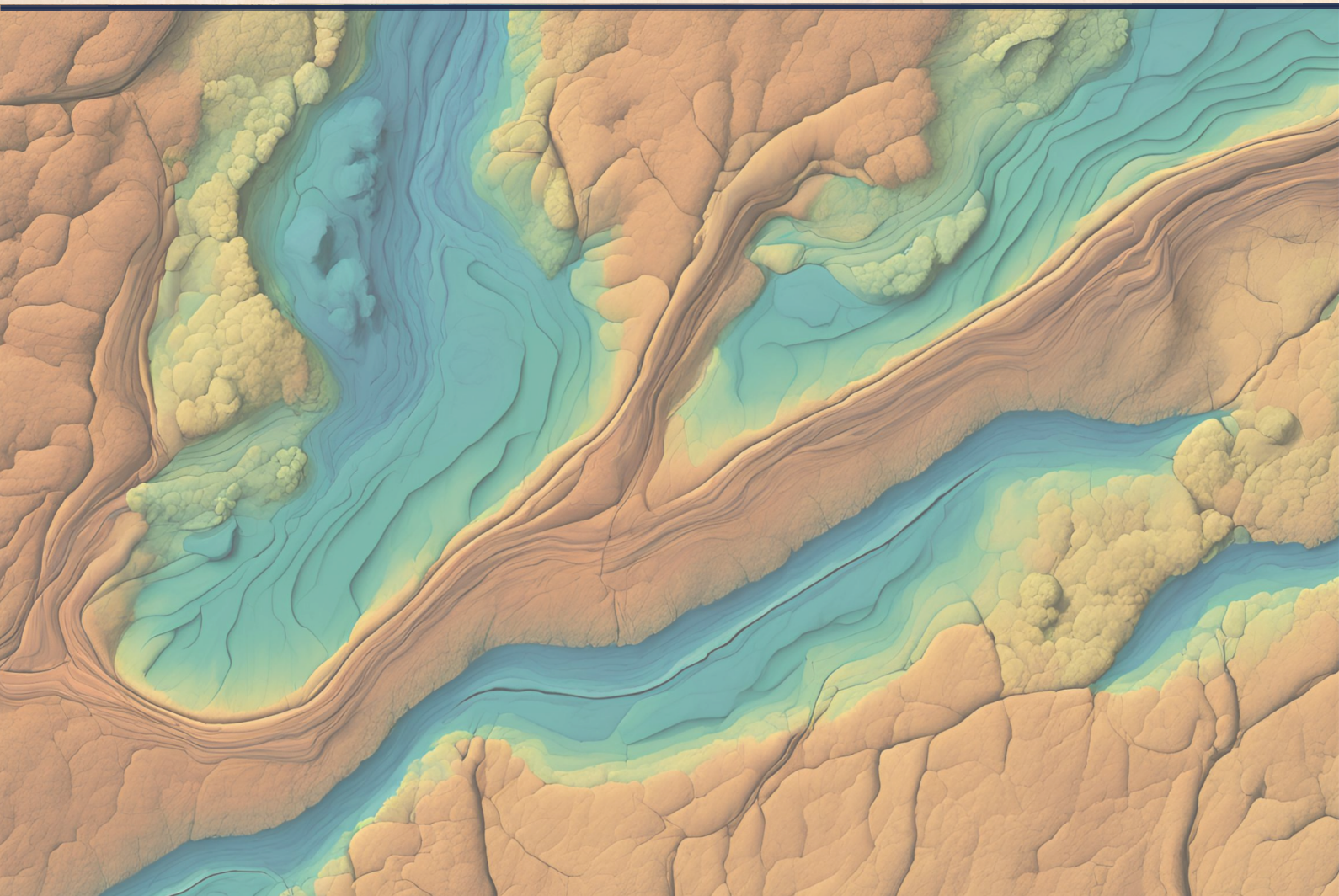
An introduction to the Future

How generalist forecasters bolster and compete with experts in any field.

An Arb Research Report

Gavin Leech and Misha Yagudin



Introduction

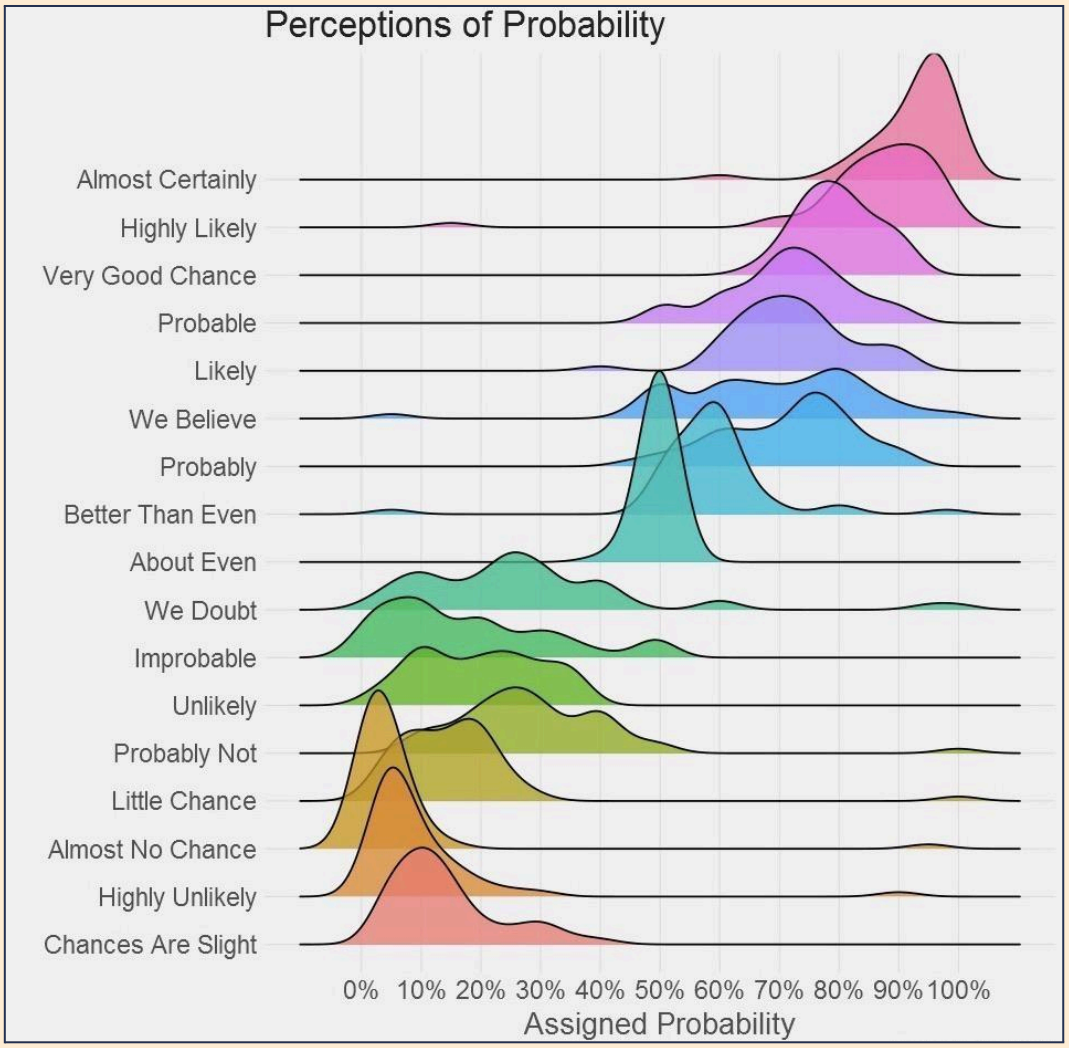
About 10 years ago a small community coalesced around the idea that certain smart nonexperts can match or beat or supplement experts at *prediction,* even on the experts' home turf. The community hopes to make society less deluded through these sorts of public probabilistic predictions. Call the method *judgement-based forecasting*, of which the famous Superforecasters are a special case – the top 2%.

Judgement-based forecasting launched with a disconcerting result about incumbent political analysts. IARPA, a US intelligence agency, ran a forecasting competition and found that outsiders could beat US intelligence professionals at their own game. Each forecaster and professional intelligence analyst put probabilities on hundreds of questions like "*Will there be a lethal confrontation involving government forces in the South China Sea or East China Sea by 31 December 2011*?". On average, the forecasters put slightly lower probabilities on things that didn't happen and slightly higher probabilities on things that did, thus predicting world events 30% better on one metric.

More recently: one of the earliest warnings about COVID as a catastrophe came from an anonymous user of Metaculus, a public forecasting platform. At the time, many public health experts and journalists were ignoring or downplaying the risk. And though that could be cherry-picked out of the thousands of predictions on the platform, Metaculus users overall also beat a panel of experts at predicting how many COVID cases and deaths the US would have had by certain dates.

Forecasters aren't magic, of course, and sometimes national intelligence merits the word. Consider the present invasion of Ukraine: beginning in early December 2021, US intelligence said the risk of invasion was "notable" and later "very high". It took Metaculus users until mid-February to give it more than 50% probability. They gave only 15% to the proposition "Will Russian troops enter Kyiv?", which came true soon after.

People often distrust claims that something is "80% likely" to happen: after all, things either happen (100%) or don't happen (0%). But people implicitly assign something like numerical probabilities for different strengths of prediction; they would never say that it was "highly unlikely" to flip a heads on a coin, and wouldn't say that a die "almost certainly" would roll a three. A classic line of work asks what *range of numbers* people mean with particular words, and produces things like this:



Making forecasts isn't new: pundits, journalists, futurists, scientists, and analysts all make claims about the future. Judgment-based forecasting is distinguished from the old kind by putting probabilities on guesses; by tracking the results of these guesses; by the use of certain debiasing tricks; and by its emphasis on *teamwork:* taking forecasts from multiple people over time, and pooling them into one overall guess.

We end up with objective measures of how good someone is at forecasting. A forecaster is well-calibrated if, when they say there's an X% chance of things happening, those things happen around X% of the time. In practice forecasts get updated after the initial guess, and scores thus reflect how *long* various levels of certainty were placed on a question. Perfect performance means always placing 100% confidence on things that do happen and 0% on things that don't. Even better is the "Brier score", which ranges from 0 (perfect) and 1 (perfectly awful), in proportion to how far a forecaster's confidence was from what actually happened. A random guess like a coin flip gets a Brier score of 0.5. (Yes, you can be worse at prediction than an inanimate object.) A world-class team of forecasters might manage a Brier of 0.12 on typical hard questions about geopolitics, implying that they're quite a bit closer to optimal performance than to a coin flip.

In her book *The Scout Mindset*, Julia Galef notes that a famous probability user, *Star Trek*'s Spock, actually has terrible calibration:

> *"There's only a very slight chance this will work," Spock warns Kirk in one episode of the original TV show, right before their plan works. The odds of survival are "less than seven thousand to one," Spock tells Kirk in another episode, shortly before they escape unharmed. The chance of finding survivors is "absolutely none," Spock declares in yet another episode, right before they discover a large colony of survivors…* ***When Spock thinks something is impossible, it happens 83 percent of the time****.*

So Kirk would do better to flip a coin than listen to his First officer.

Averaging a group of forecasters often does better than asking one person, since they know different things and cancel out each others' biases. This is the famous 'wisdom of crowds': most people asked to guess how many jellybeans are in a jar do poorly, but the average guess is remarkably close. Rewarding or punishing forecasters based on their performance does even better than that. Hence prediction markets, where participants risk their own funds on the strength of their convictions. The Good Judgment Project, an academic team working on a US national intelligence project, was the breakthrough for this kind of forecasting. Their big contribution was that teaming up and advanced aggregation

matters, and lets you do better still than markets.

We now take a closer look at the recent history of upstart generalists, at *how* the best forecasters perform so well, and then try to explain why this niche and nerdy pursuit matters for everyone.

The upstarts: data scientists, markets, and forecasters

Forecasters were not the first 21st century shock to the unique status of experts: that was the analysts, wonks, and quants of books like *Moneyball* (2003) and *Super Crunchers* (2007) and sites like *FiveThirtyEight* (2008). They used data to outdo pundits in sports, politics, and other data-rich domains. This is model-based, rather than judgment-based in our sense: their predictions come from statistical models where the human takes themselves out of the loop after setting up the model. But the algorithmic approach misleads when the data are lacking, as they so often are. This is where judgment-based forecasting is most useful.

Michael Lewis, who covered the analytical approach to baseball, notes that there's an oddly strong resistance to this kind of data science:

> *When the data-driven approach… did not lead to immediate success — and occasionally even when it did — it was open to attack in a way that the old approach to decision-making was not… "We have perhaps overly relied on numbers…" said [Red Sox] owner John Henry…*
>
> *whatever it is in the human psyche — this hunger for an expert who knows things with certainty, even when certainty is not possible — has a talent for hanging around. It's like a movie monster that's meant to have been killed but is somehow always alive for the final act.*

Other industries are based on pooling judgments at scale, too. Stock markets produce forecasts as a byproduct: people buy stocks predicting that the price will rise, or short them predicting it will fall. This can tell us a surprising amount about the world, indirectly – witness the crash in Russian stocks a full week before the Ukraine invasion. And there's betting on sports, which is hugely popular, and produces odds that reflect what bettors willing to put money on the line think will happen.

Besides asset prices and sports, we have the general prediction markets, which allow us to bet against each other about anything legal. You could see them as decentralised bookmakers – where getting accurate odds is the whole point. They've been running since the eighties, with the most popular covering political events. But the current prediction markets are tiny for everything except the biggest tickets, like the US presidential election. (In fact, US election markets have run since the 1800s.) But there's tens of thousands of times more money in sports and election betting than all other events combined, and general markets have been

killed off several times by regulators. Lately some new markets found purchase in the crypto world, hiding there from the feds. An ingenious idea is to shunt a market's profits off to charity, to avoid gambling statutes while leaving us the precious predictive signal.

But for now we don't have the sort of large, liquid prediction markets that we'd need to really get the wisdom of the crowd for questions like, say, whether Germany will re-open any nuclear power plants by 2030. If we can't bet on things like this, we need some group to predict them.

Traditionally we looked to pundits and academics. But consider their failures: nearly no political scientist or Kremlinologist predicted the fall of the Soviet Union; pundits routinely spout convenient or eye-catching nonsense.

The new approach to forecasting started with Philip Tetlock, whose 2005 book *Expert Political Judgment* quantified the bullshit in area studies, geopolitics and allied professions. Experts were often shockingly inaccurate and miscalibrated about the *specific* geographical region they had spent years studying. Worse: in general the more famous experts were, the less accurate. In messy fields like the social sciences, we apparently can't presume that *any* given expert is correct or calibrated, though there are efforts to forecast there too.

And then we have judgment-based forecasting. How do successful forecasters operate? It's mostly not about modelling or stats, though top forecasters tend to be comfortable with stats. And the new forecasting goes beyond prediction markets by foregrounding teams of forecasters and behavioural science.

The same people often forecast in markets and in pooled forecaster teams. The difference is teamwork and how cleverly their predictions are combined.

The hard evidence

We wanted to figure out whether the forecaster advantage has held up, so we looked at a decade of studies comparing top forecasters and experts. There was less research than we expected. Some results have been misinterpreted, too: the glorious number from above, that aggregated forecasters were "30% better" than intelligence analysts, turned out to be an unfair comparison – the intelligence analysts were combined using a prediction market, while the forecasters' work was combined using a better technique, extremising and weighting by past performance.

Still, a few good studies suggest that top forecasters are at least on par with self-selected experts – and aggregated forecasts usually beat an expert forecasting alone. One of the better studies had experts and forecasters predict disease spread and found that the two groups did about as well as each other. Mixing forecasters and experts into one team also looks promising, given their presumably complementary skills. That's obvious in hindsight, but people didn't try it until quite recently. Note that we don't want just any experts to join the teams; we want unusually flexible ones with a strong grasp on probability.

We can say some tentative things about different domains: In one great study, generalists at least matched experts at predicting geopolitical events, and probably did better. The forecasting service Hypermind did remarkably well at the 2014 US midterm elections, beating major newspapers and FiveThirtyEight. Forecasters do very well at COVID questions – maybe 3% to 10% better, though this was mostly overworked experts at the busiest time of their lives, and the top performers were public health pros with forecasting training.

Overall, it looks like methods for aggregating predictions are great, and the top generalists are capable – but combining them with savvy experts might be best.

The rest of this piece is more speculative. Some of the numbers are made up — sorry, I mean judgmentally forecasted.

What explains forecasters matching experts?

It's impressive that certain outsiders can match experts on their own turf. How can this be true? We already mentioned teamwork, aggregation methods, and the bag of cognitive tricks. What else? True expert forecasting has never been tried? Start with the default boring explanation: have we been comparing the best forecasters to the average expert?

Maybe. Experts are busy. And the better, the busier, since they'll tend to be in higher demand. So when you look for volunteer experts with time to spare, you'll get a nonrandom sample. Since they have a day job already, this sample of experts might update their forecasts less often than the forecasters, and frequent updates are one key ingredient of forecasting success. But by the same token, their volunteering for the project reveals that they care more about the task than the average expert — and caring about being accurate also matters a great deal.

Nor have we managed to show generalist forecasting at its full potential: almost all prediction markets are missing one of the key ingredients: real money stakes, many users competing, making it easy to make forecasts, *and* making it easy to ask forecasters new questions. Against this, Tetlock and Mellors suggest that the existing top teams might be as good as humans get:

> *No one ever expected a deterministic world in which Brier scores of zero were possible—and many observers were surprised that Brier scores could be pushed as far down as they were, falling as low as .12 to .14 for the best polling algorithms... we may be reaching the point of irreducible uncertainty—and IARPA should be skeptical of future investments in geopolitical forecasting.*

Overall, we think it's possible to get even better forecasters *and* better experts, if we manage to get the right incentives and infrastructure, such as larger prediction markets paying out real money, or media editors paying more attention to track records of their columnists and guest writers. Wouldn't you rather that op-ed come from someone who's more often right than not?

1) Practice

Most people don't deal in explicit probabilities and are not trying to improve their forecasts. They reject bets when their beliefs are challenged. Very few professionals deal with explicit probabilities and unambiguous outcomes in their normal work. So maybe they just haven't learned this precise skill. For instance, American doctors massively overestimate the probability of certain diseases given test results, forgetting that many diagnostic tests give high rates of false positives for relatively uncommon illnesses. The probability of a woman having breast cancer after getting a positive mammogram is 6%, because of false positives, while this sample of American doctors said it was 50%. But there are signs that they are getting better – they used to say it was 75%! So we might be able to improve day-to-day "forecasting" like this just by making people more aware of the basic tools they can use to think about the probabilities they deal with. The best Metaculus forecaster went from no experience to winning second prize at a huge IARPA tournament in one and a half years.

2) Lack of preconceptions and misleading narratives

Maybe pros come to view everything from one narrow viewpoint which overlooks crucial information. Or maybe their detailed knowledge is itself an impediment. Consider Kahneman and Tversky's idea of the *outside view*, which uses the odd power of generic examples to remove the biases and preconceptions people often bring to their analyses, and avoid overblowing minor details. An expert predicting the outcome of a land war against Freedonia can take the inside view, drilling down into all the details: what the opinion polls say, which generals are in charge, and which weapon systems each side has acquired.

Experts know more such details, and thus have more freedom to tell themselves a complicated persuasive story and miss the wood for the trees. The outside view instead asks simple questions like "what fraction of land wars in this region are successful?" or "how often does Sylvania win wars?" and in practice this is surprisingly effective. Tetlock again:

> *Unpacking scenarios encourages people to adopt an inside view: to immerse themselves in each case and judge the plausibility of pathways to outcomes by drawing on their detailed case-specific knowledge of the forces at work… [unlike] if they had stepped back from the details of individual cases and grouped them into summary categories (base rates)…*

3) Decorum

An expert's public statements might not exactly match what they think, due to extra commitments, professional decorum, or social desirability. A forecaster might instead give a blunt opinion, since they have fewer professional relationships to maintain. Returning to our doctors from above, it's easy to see why they might overestimate the likelihood of disease: they practice "defensive medicine" to protect themselves against malpractice suits. Generalists often don't make such decisions, and are thus insulated from responsibility. A related angle is simple selection: if you didn't firmly believe that your field's methods are good, you wouldn't stay in the field.

4) Intelligence

Are top forecasters smarter than experts? Tetlock:

> *Regular forecasters scored higher on intelligence and knowledge tests than about 70% of the population. Superforecasters did better, placing higher than about 80% of the population.*

But the intelligence section of the Forecasting Aptitude Inventory he used was extremely short, and tests this short can't distinguish high performance from *very* high performance. It's also unfashionable to administer intelligence tests to adults, so we don't know how this compares to the relevant experts in say politics and public health.

Top forecasters excel on one measure: cognitive reflection, or the ability to override your gut reaction and decide when to look closer. Great forecasters know they won't always come up with the right answer immediately and actively revise their initial views.

5) Good incentives and public track records

Forecasters can now get paid; a number of them have made a career off their verified reputations. So there's a system in place for making them care about each specific claim they make. It seems rarer for academics to be held up on specific claims—and when they are it's often in private peer review.

6) Foxy personalities

One of the big constructs in the Tetlock team's research is a spectrum from "foxes" (pluralists, bottom-up thinkers) to "hedgehogs" (monists, top-down thinkers). Foxes do better because they're more willing to switch approaches and let go of a pet theory. We might expect experts to be more hedgehoggish and remain attached to a bad theory that they've worked on for a longer period of their lives.

7) Expertise is better for facts, but forecasting can beat it under uncertainty.

Experts sometimes lay claim to a large area of discourse. But often expertise is useful for a subset of questions—"what species is this?","how infectious is this agent?", or "why do the Saudis hate the Iranians?" where the question requires specific and often rare knowledge to answer, rather than a question like "how many worldwide infections will there be from this novel pathogen?", where your handling of uncertainty helps more than your knowledge.

Experts are also forced to specialise, to cover a relatively small part of their field. In search of expert feedback, you might ask "a computer scientist" questions about computing — but cybersecurity people do extremely different work from the programming language theorists, who do extremely different work from the human-computer interaction people, who do completely different stuff from the quantum computing people. Their expertise would give them little extra help in each other's fields – in contrast to forecasting, where the whole point is to be able to answer questions in very different domains. The top Metaculus forecaster has excelled at questions on ebola, gold prices, Polish elections, North Korean missiles, and Starcraft, among other things.

Studying a slightly different question — what predicts an individual forecaster's success *before* aggregation — the Tetlock team's model implies a 55% contribution to predictive success from teamwork and effort, 23% from intelligence and fox mindset, 10% from forecaster training, and 12% from your degree of domain knowledge. These numbers come from the coefficients of their structural equation model, a way of taking into account complicated causal relationships and coming out with good estimates of what contributes what. This omits the likely benefits of aggregation mentioned previously.

Where doesn't forecasting work?

There have been plenty of flubs. In 2021, a large group of crypto bugs attempted to buy a copy of the US Constitution at Sotheby's. Just before resolution – that is, when all the information was in – the Polymarket entry on the auction was trading at 700-to-1 in the wrong direction because the market overreacted to some fake news. The 2016 Trump upset was barely less surprising for forecasters: Metaculus implicitly put 10% on Trump winning.

Those are just random mistakes out of thousands of good community forecasts. But is there anywhere we should expect forecasting to be systematically below par? This hasn't been formally studied, but we can guess:

Big boy markets. Asset markets are the original wisdom of crowds, using the most powerful incentives in the world. It seems possible that the best forecasters don't even go by that name: maybe the professional quants and traders already predicting market prices for us would otherwise claim the top spot. But are quants weird enough to predict unusual events? We're not sure. On one occasion forecasters did react slightly faster than the market to a COVID shock.

The story is complicated though. Top finance people should be better at forecasting than superforecasters because their domain overlaps with forecasting and because the financial incentives let firms select the best people in the world. Most people with the opportunity to earn a hundred times more money will do so.

We speculate that there's a second selection effect: to forecast well, you need spare time in which to update your predictions. But this need for spare time rules out most quants and analysts.

Arcane disciplines. Extremely technical domains will also give forecasters trouble — anywhere where normal intuitions are foiled by the sheer number of moving parts. This obviously applies to open research problems in mathematics, where just understanding the problem statement is often beyond outsiders. Outsiders probably can't add much here, beyond looking at the average time between someone posing a problem and someone coming up with a solution.

Places with secrets. We start a fresh prediction using a base rate: how often did this kind of thing happen in the past? If there's missing or bad data from survivorship bias, confidentiality, or active poisoning, this base rate can be off by orders of magnitudes and ruin our whole process. So in principle, insiders who know how the process is being censored could dominate. But then thinking about missing data is just another key

forecasting skill.

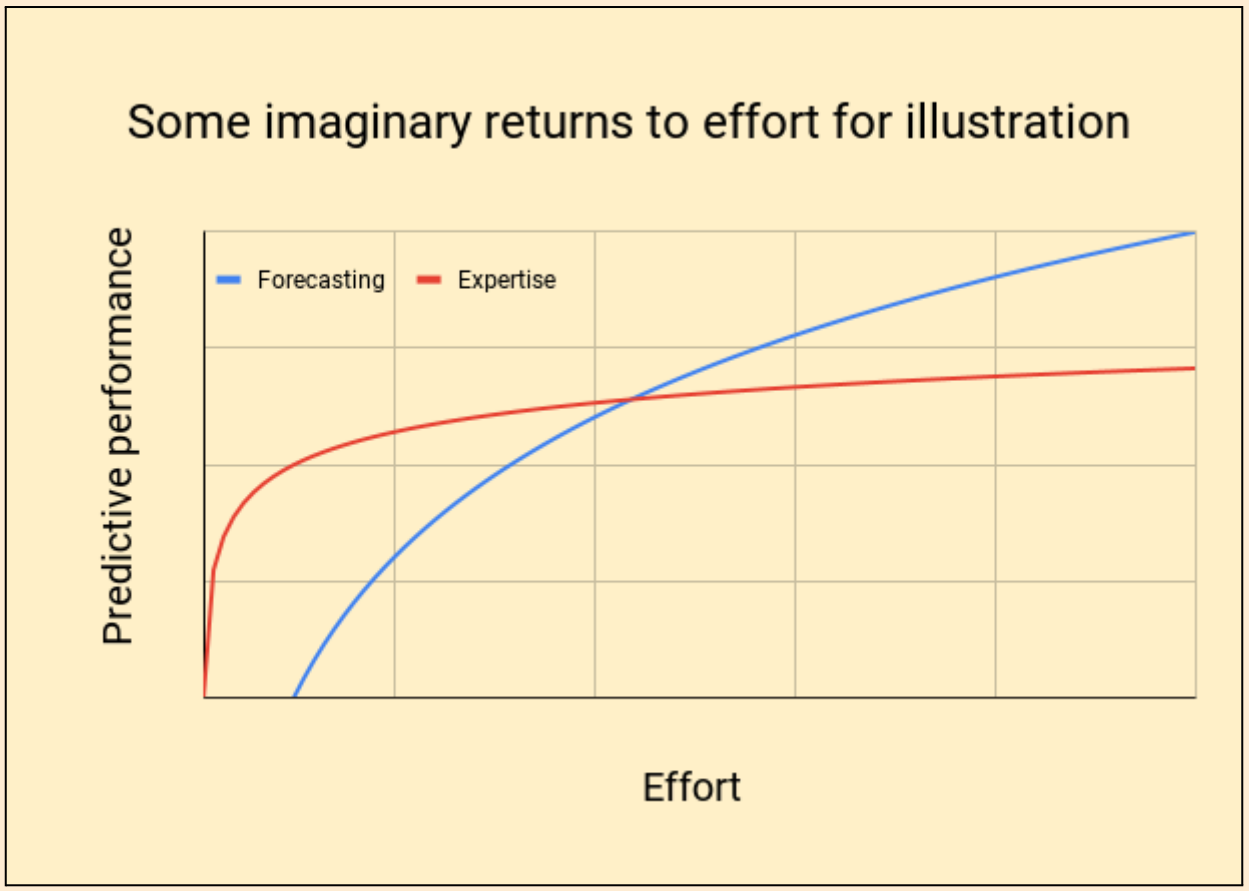
We shouldn't exaggerate this one. People tend to assume that there's a really high return on expertise – that spending years on something, making one topic your focus must make you good at predicting it, the natural person to defer to. But as we saw, Tetlock's earlier work shows that this deferral isn't warranted in politics.

People like to assume the world is full of useful secrets and omnicompetent deep states — but as far as we can tell from studies that pit national intelligence people against generalists, these are rare.

Backtracking

The above assumes a clean distinction between "forecasters" (generalists with very good reasoning ability, practice with probabilities and very little domain knowledge) and "experts" (specialists with some reasoning ability and a great deal of domain knowledge). But this isn't clean at all, and there are many examples of overlap. For instance, the top Hypermind forecasters in Sell et al, which tried to predict the path of the COVID pandemic, were public health professionals. And that's before we consider teams mixing forecasters and experts. Better to say that, after a certain point, the benefits of additional expertise are marginal for many important kinds of estimates.

So don't ask "how good are forecasters compared to experts"; instead, ask "how much performance does more expertise get you in this field?" and "how much performance does more forecasting skill get you in this field?". Here's a made-up graph to show you what we mean:



Of the variables studied by Tetlock's team, the degree of domain knowledge was associated with only a middling amount of forecaster performance. And we know the rough shape in geopolitics already, from Expert Political Judgment:

> *[T]he consistent performance parity between experts and dilettantes... suggests that radical skeptics are right that we reach the point of diminishing marginal predictive returns for knowledge disconcertingly quickly [in politics].*

The upshot

There'd be a lot to like about forecasting even if it didn't beat experts. For instance, it lets us choose to listen to better people. Scott Adams is a popular pundit with a poor track record.

Similarly, Metaculus' Public Figure Predictions project looks at the loudmouths that dominate our public square and lets us check whether they *should* dominate.

Then there's a connection between forecasting and *virtue*. One virtue of a scientific theory is whether it's falsifiable: does the theory admit the possibility of disproving it? But this isn't only good for science! "Is it falsifiable?" is equivalent to "would you bet on it in a way you couldn't wriggle out of?", and pretty close to "is your opinion actually adding anything to the discussion?".

Another huge benefit is that forecasters are just more common than experts in any particular area, so on the assumption that our experts have less time to make and update forecasts, the generalists can step in and massively expand the range and volume of what gets forecast.
Forecasters have also been more willing to be publicly tested, which makes it easier to select the best performers there – we don't need to interview someone, we can just check their scores in seconds.

More grandly: some people look for ways to improve society's *epistemics* — how accurate we all are, how proportionate our confidence is, how robust we are to manipulation and bullshit. We want to fix the incentives of everything, to make information flows which are too hard to corrupt. Better forecasting is one of the main hopes for this. It scales better than hallowed philosophical instruction, anyway.

One thing the community has given us already is a new way to tell the news. The writer Scott Alexander has been providing weekly Ukraine updates as shifts in Metaculus' predicted probability of particular events. This summarises the opinions of hundreds of clever people, weighs them by how right they've been before, and hands it to you in the densest form. Maybe this is the journalism of the future.

What does all this imply for you, the reader? Well, sometimes outsiders match or beat the insiders: so take your own ideas seriously, keep track of your exact positions, and *vote your credences*. If others were right, update towards them. Don't just "do your own research"; *find out if you should.* Merge your guess with others, including experts. Join in, or support those who join in. You might catch something everyone missed; you might be the first in the world to notice.

Gavin Leech and Misha Yagudin run Arb, a research consultancy that can't seem to stop working on forecasting. They're hiring! Catch them on Twitter here.