Nesta...
WORKING PAPER SERIES

Nesta Working Paper No. 15/02

Prediction as sport: the promise of prediction polls and markets

Jeremy Kingsley

Prediction as sport: the promise of prediction polls and markets

Jeremy Kingsley

Nesta Working Paper 15/02
February 2015

www.nesta.org.uk/wp15-02

Abstract

You can in certain circumstances combine the moderately accurate judgments, or forecasts, of many individual people to produce a collective judgment this is more accurate – one frequently more reliable than any lone expert. Two ways of doing this, specifically for forecasting future events, are the focus of this working paper: prediction polls, in which many individuals are asked independently to assign a numerical probability to a future event and those judgments are aggregated; and prediction markets, in which individuals bet against each other about the likelihood of events, and the market price is interpreted as a probability. Much evidence suggests that these tools produce predictions that are very accurate, and can improve our ability to forecast a wide range of phenomena – be it geopolitical and macroeconomic events, scientific breakthroughs or the success of an emerging technology, or a company's sales targets. Evidence also suggests that participation in such forecasting tournaments improves, with training and regular practice forecasting events, improves individuals' ability to forecast dramatically. The aim of this paper is to explain the principles behind these tools, uncover in what circumstances they work best, and explore how such ideas could be used to make better forecasts that in turn inform better decisions, not just in theory but in practice.

Keywords: Forecasting, prediction markets, collective wisdom, polling

For their time, conversation and feedback my particular thanks to Pavel Atanasov at the Good Judgment Project, as well as to Jessica Bland, Bo Cowgill, Robin Hanson, Hubertus Hofkirchner, Regina Joseph, David Rothschild and Justin Wolfers.
Corresponding author: Jeremy Kingsley, Nesta, 1 Plough Place, London, EC4A 1DE, jeremykingsley@gmail.com.

Table of contents

Prediction as sport: the promise of prediction markets and polls

Off the back of 18 years cataloguing tens of thousands of experts' forecasts of political events, the psychologist Philip Tetlock roundly denounced the predictive powers of the political scientists, policy wonks and media pundits whose predictions he tracked as little better than chance.[1]

But rather than going on to dismiss long-term political forecasting as impossible, Tetlock put his efforts into working out how to do better. The results have been more encouraging than anyone, Tetlock included, could have expected.

Tetlock now co-leads the Good Judgment Project, a large-scale research programme funded by IARPA, the US intelligence agency. Good Judgment was one of five forecasting teams funded in IARPA's 'ACE' programme.* ACE, run out of the agency's 'Office for Anticipating Surprise', was established to:

> dramatically enhance the accuracy, precision, and timeliness of intelligence forecasts for a broad range of event types, through the development of advanced techniques that elicit, weight, and combine the judgments of many intelligence analysts.

Typically, as individuals, we are not very good at forecasting – as Tetlock showed. But as has been amply demonstrated and discussed,[2] you can in certain circumstances combine the moderately accurate judgments (in this case forecasts) of many to produce a much more accurate collective judgment.

Two ways of doing so are the focus of this working paper: prediction polls, asking lots of people and taking an average of what they say; and prediction markets, in which people bet against each other about the likelihood of future events. Evidence suggests that these tools produce predictions that are much more accurate than other methods, and can improve our ability to forecast a wide range of phenomena – be it

* Aggregative Contingent Estimation, <http://www.iarpa.gov/index.php/research-programs/ace>. After two years, the forecasts of Good Judgment, the sole academic team (based between the University of Pennsylvania and Berkeley), were found to be much more accurate than the other teams. So much so that IARPA cut its funding for the other teams and reallocated resources to concentrate on Good Judgment.

geopolitical or macroeconomic events, technological disruption or a company's sales targets – and in turn help us make better decisions.

A recent comparison of public US midterm election predictions found that Hypermind's prediction market, a venture similar to Good Judgment that puts 'elite' forecasters in elite markets, far out-performed many statistical models – including FiveThirtyEight's.[3] As such, for such phenomena as elections, collective human judgment is better than individual experts, and data-driven methods too.

What follows is a brief investigation into these tools for the purposes of better understanding them, and the ideas that underpin them – including collective wisdom, framing questions about the future, and training individuals to be better forecasters. When do and don't these techniques work? And can we use such tools, or the insights gleaned from academic research into them, to get a better handle on the future?

*

Good Judgment has invited thousands of members of the ordinary public[†] to predict events that matter to the intelligence community, such as the fall of a regime, an election outcome or the likelihood of a country defaulting on its debt. It has experimented with the many ways you can take those individual forecasts and aggregate them, with the different ways individuals can interact (in group discussion, or trading against one another in a market), and with ways of making individuals' raw forecasts better – such as through training. Questions vary in their time to resolve, some settling as soon as within a week; a year is the maximum.[‡] The tournament is now in its fourth and final year, and will report on its findings after the project ends in mid-2015.

The bulk of the work at Good Judgment has explored ways of collecting individual judgments together and producing crowd forecasts that are more reliable than lone

[†] In practice, forecasters are overwhelmingly male and more likely to have a postgraduate degree than not – see the note on diversity (§1). See also: <https://goodjudgmentproject.com/blog/?p=242>
[‡] This is largely a constraint of the research format, which runs as an annual tournament. There are other difficulties of running long-term polls and markets, however: principally that participants lack incentive if the pay-off is so distant.

experts. But a significant finding so far – and the one that has received the most press – has been that some people really are good at foresight (significantly better than random guessing), and individual forecasting skills can be honed. It was discovering these 'superforecasters', and putting them in forecasting teams, that set Good Judgment apart from the other programmes originally funded by ACE. Its polls and markets performed better, said Pavel Atanasov, a post-doc fellow at Good Judgment: "but the killer app was the superforecasters."[§]

As such, understanding how these tools work helps us work out where to put them to good use, as well has how to be better forecasters ourselves – and both aspects will be explored here.

[§] Other projects, such as Hypermind <http://www.hypermind.com>, have developed similar approaches.

§1

Mechanics: polls and markets

Prediction polls

In a *prediction poll*, a group of individuals are asked to assign a numerical probability to what they regard as the likelihood of a future event. Those many independent judgments, whether a handful or several thousand, are then aggregated to produce an overall crowd forecast – at its simplest by taking the average.

This is a straightforward adoption of the idea that wisdom can be found in crowds. The general idea is well described by Condorcet's Jury Theorem: if you gather the independent judgments of a group of people, if each is individually more likely to be right than wrong, then as the size of the group increases the likelihood that the collective judgment (the simple average) of the group is correct tends towards 100%. Even if each person is individually only right 51% of the time, you will soon descend on a highly reliable judgment with a sufficiently large group. Or put it another way: it implies that a large group of ordinary intelligent people can be as accurate as a small group of experts.

But more sophisticated aggregation than a simple ask-and-average operation can greatly improve the crowd's wisdom. For the polling mechanism to produce more accurate forecasts (as determined after the fact), the statistical algorithm that takes individual forecasts and produces a collective one can be tweaked, and individuals' judgments weighted.

First, weighting the forecasts of individuals who are believed to be more reliable (based on their track record, say, or even a ranking of their cognitive abilities – see §4) can significantly improve the collective forecast. If forecasts are made and revised over time, then discounting older forecasts and promoting fresher ones, which will take into account the latest information available, helps considerably too.

Second, there may be consistent biases in individual and collective judgments that can be corrected for. "Our aggregation algorithms mostly stay away from de-biasing,

because we're truing to invest as much effort into making the raw forecasts better," said Pavel Atanasov about Good Judgment's prediction polls. "With that said, there are some patterns of aggregate predictions that are biased, even if the individual is not."

For instance, the average of many judgments will tend to be under-confident – too close to 50% (as opposed to the extremes, 0 and 100%).[**] To counteract this, the numbers are typically 'extremised': pushed away from 50%.[4] "A formula takes the raw prediction and pushes it towards the extremes," explains Atanasov. This happens after the mean is taken, not before. Doing so improves the accuracy of the crowd prediction at Good Judgment by about 10%, he says.

Much of Good Judgment's research is into discovering what factors like these matter most, and what kind of weighting and statistical massaging can produce the best predictions. Such algorithmic tweaking and trial and error is valuable for working out how to get the most accurate forecasts from a group of individuals, though it adds a layer of complication – any algorithm that makes assumptions about individuals' biases may not generalise to all groups, and remain true over time. For this reason and others, many promote the idea of prediction markets – which are believed to dispel the influence of bias and error automatically, simply by the power of market forces.

[**] Individually, humans tend to be overconfident about unlikely events and under-confident about more likely events – the 'favourite-longshot' bias (this is predictable and regular, and so can be corrected for; see also the end of §2). A group will also be under-confident even when the individuals within are not: that is because, first, probability is bounded (individuals can only give between 0 or 100%), and second that the aggregate of many people has more information than the forecast of just one person – but not necessarily more confidence.

Prediction markets

In a *prediction market*, people bet against each other on the likelihood of events. Whilst polls involve asking people directly, these markets yield predictions indirectly through individuals' behaviour: following their self-interest, traders will move a market price up and down, which can be interpreted as a collectively-determined probability.

Prediction markets are motivated by the idea that financial markets reflect different predictions about the future returns of an asset, settling on a point of agreement – the price – that reveals a collective judgment. Markets are interesting because, though sometimes very wrong, financial markets are often smarter than the individuals within. Investors rarely out-perform the market over the long-term.

This idea of markets as an aggregation mechanism follows Hayek.[5] He, in the particular context of an argument against centrally planned economies, put it that the knowledge we require in the smooth running of an economy exists only as the dispersed bits of knowledge which separate individuals posses: the market mechanism effectively reflects information partially held by many, as if "a single mind possessing all the information".

As such markets can be powerful tools for eliciting and aggregating information and expertise that is dispersed amongst large groups. 'Prediction markets' (or information markets) are established solely for this quality, with participants trading contracts relating to a specific future event – say that Hilary Clinton will be elected president in 2016 – that yield payments if the event happens. If the expiry price is £1, say, then a prediction market contract trading at £0.40 can be interpreted as the crowd regarding the event as 40% likely. Anyone who thinks it's more likely will ('rationally') think 40p is cheap, and buy it up – and vice-versa.

Rational trading behaviour expects that the value of a contract in the eyes of a trader at a given time to be a product of their confidence in the event transpiring (a subjective probability assessment) and the value of the expiry price of a contract to

that trader.[††] (It's worth pausing to note the rational standard this expects of prediction market traders – qualities over and above their ability as forecasters.)

One immediate attraction of a market is that little of the special statistical aggregation that is involved in polling, as described above, is required. Another is that, because traders 'put their money where their mouth is', volumes of trade correspond roughly with confidence, a measure missing in prediction polls.

"The problem with polls is you're not giving people much incentive to focus on the questions they know better," says Robin Hanson, an economist at George Mason and a pioneering architect and advocate of prediction markets. "And you're not giving them much of an incentive to try since the pay-off isn't dependent on how well they do."

More, the idea goes, any biases that may be at work that may have to be corrected for in other situations should be apparent in the price, and so won't last long – it presents a profit opportunity to someone. "You have an incentive to look at the distribution of prices, notice any biases, and fix them," says Hanson. "There's a meta feature of self-correction."

Prediction markets have received a lot of support and interest in recent years.[6] Though the most well-known prediction market, InTrade, collapsed spectacularly in 2013, having faced years of regulatory trouble, there are increasingly many real-world markets letting people speculate on politics, sport, economics and entertainment. Prediction markets have been established in major corporates including General Electric, Google, Intel and Microsoft, and by the likes of the US Department of Defense, on a wide range of topics from sales targets to security threats. Their predictive record remains very good. Whilst polling in the run up to the Scottish referendum was shown to be poor, for example, prediction markets (particularly

[††] Many factors determine how many contracts someone is willing to buy or sell, including how much their perceived value differs from the market price, how much money and contracts they have, the opportunity cost of trading one outcome's contracts rather than contracts for other outcomes that might be listed concurrently (see Servan-Schreiber 2012, pp22-3). All these factors may play a role, and different combinations of them may yield the same trading behaviour, making it difficult to interpret individual trades. More, particularly in real-world, real-money markets, the trade may be merely speculative, based on short-term movements on the price.

Betfair, the world's largest) fared much better, and the same has been true of US elections.

"Prediction markets have pretty consistently shown that they produce as accurate or more accurate forecasts than other methods, on a wide range of topics," says Hanson. "Consistently, when there are comparisons of prediction market accuracy against other institutions or mechanisms making forecasts at the same time, with remotely similar levels of resources and participation, then prediction markets either give pretty much the same answer or they give substantially better answers."

Commonalities

There is nothing inherently wise about a crowd, and there are certain conditions that should be satisfied in order for a group to land on a reliable collective judgment, rather than one that accentuates bias, misinformation and results in group-think. Typically, you need a diverse, decentralised group, and for individuals' judgments to be independent. At Good Judgment, different conditions have been experimented with to gauge what impact certain design features can have. Two key ideas are independence and diversity.

Independence

Agents' independence is often espoused as a vital precondition of collective wisdom.[7] When there is a lack of independence, there can be a tendency towards 'groupthink': common knowledge gets emphasised, whilst privileged information doesn't come to the fore. One reason is an information cascade, in which participants discount their own knowledge: when people speak in turn, what's said first can have a huge impact and the next person will adjust according to the consensus. As one person adjusts their judgment to the previous person's, the effect becomes more pronounced with each subsequent person in line – and strength of opinion grows.[8] Groups of people, particularly in organisations, are also prone to 'reputational' cascades, in which participants silence themselves to avoid blame, criticism, opprobrium or other reputational sanctions.

The advantage of decentralised polls and markets is that you avoid these pitfalls of deliberating groups. Nonetheless, the Good Judgment Project has experimented with polls in which forecasters deliberate in teams of 15. In those groups, participants

share information, debate and discuss their reasoning (through a digital forum) and then they are judged independently and as a team. The results so far suggest that in this dedicated forecasting environment, in which participants are wary of groupthink, the gains of groups sharing information exceed the losses (see §4: teaming). “We are hyperaware of our biases,” says Regina Joseph, a futures consultant and superforecaster in one of the best-performing teams in the tournament. “Sometimes there’s a lot of uncertainty around a question, and that’s when it’s useful for me to scan and communicate with my teammates. We are hyperaware of groupthink, we talk about it all the time.”

A different dynamic with respect to independence plays out in prediction markets. In one sense, forecasters are entirely independent in their trading, not deliberating with others. But they do see each other’s trades, and the price acts as a signal as to what the market is thinking.

In financial markets, paying more attention to the price and its movements over time, than underlying information, is for many the point of trading – particularly when there’s great uncertainty. Many participants are engaged in a ‘game’ Keynes famously described akin to a beauty contest in which people are anticipating what average opinion expects the average opinion to be. Many have shown that when traders pay more attention to the price and the beliefs of others, than their own knowledge, then information can become ‘trapped’ – with herd behaviour taking over from efficient market aggregation.[9] (This is remarkably similar to an information cascade.)

But in another respect, the price signal can be extremely useful, acting as an anchor to your own knowledge, suggests David Rothschild, an economist at Microsoft Research and founder of the company’s Prediction Lab. Say you know that an election candidate had a good day and so their probability should increase. But you don’t know what their baseline probability of winning is already. In this case, the market price can guide you. Without it, privileged information – that someone’s odds increased, but without any idea as to the overall likelihood – can’t be added into the mix.

Diversity

The diversity of the forecasting population is also heralded as a key precondition to successful collective wisdom: the more diverse the group, the greater the pool of the knowledge and the more cognitive perspectives are brought to bear on a particular question. The greater the diversity, "the more complimentary bits of truth can be combined, while the extra biases still get canceled."[10]

For some, the key to collective wisdom is this diversity. Scale, individual expertise and independence are all secondary.[11]

Good Judgment's pool of forecasters is not particularly diverse, demographically at least. Its forecasters are overwhelmingly male and well-educated. They do participate however from all over the world. Good Judgment has looked at diversity in their experiments, to see whether more diverse groups do better. "Frankly we haven't found any support for the hypothesis that any kind of measurable diversity seems to help," says Atanasov.

§2

Relative merits

One of the most important things at work in these forecasting methods is holding forecasters to precise judgments – turning vague, subjective pronouncements about the future into specific, numerical forecasts. That makes it easier to score individuals' forecasting ability, to compare several individuals' forecasts of the same event, and to make decisions based on those forecasts.

But assigning numerical probabilities to highly specific events is obviously much harder to do. It's relatively easy to predict the future without a timeline: we're confident that we'll get driverless cars 'soon', but it's much harder to say exactly when that's going to become mainstream – and in what way.[12]

Asking questions that resolve exactly, but are still useful to guide decision-making, is not trivial (see §3). But in order to compare forecasting abilities, and act on the results, such rigour is vital.

Relative merits

The Good Judgment Project will report on the results of its experiments after the project ends later this year (2015). The researchers have a variety of forecasting conditions, including independent polls, team polls and two kinds of prediction market (one operating by continuous double auction, like the stock market, the other with a logarithmic market scoring rule, which improves liquidity issues[⌗⌗]).

In the tournament's second and third years, the researchers ran a randomised controlled trial of the different conditions. They found that independent polls are about as good as the markets, whilst, with the proper aggregation, team polls outperform both significantly (12 to 20% more accurate).

Prediction markets tend to be comparatively worse than polls at long-term forecasts, becoming more accurate towards the close of the question.[13] "Right at the start of

[⌗⌗] In short, different trading mechanisms can allow for more trades to be matched, so improving the accuracy of the price.

very long questions, ten months from resolution, markets appear to be quite a bit more noisy and less accurate," says Atanasov. "That's where the advantage of polls is coming."

One reason may be that people don't care as much to bet on long-term events. They can turn around their money more quickly with short-term contracts. Atanasov says that the more accurate and involved forecasters (the 5% of traders who make up half of trading activity) tend to prefer questions that resolve sooner.

Another shortcoming of prediction markets is that they, like ordinary betting markets, are known to fall foul of the 'favourite-longshot bias': in which the chances of the underdog are overrated. This kind of bias can, however, be corrected for – since it's predictable and consistent. For example, Nate Silver's FiveThirtyEight 2008 election forecasts were found to be slightly better than those of InTrade, the public prediction market – but InTrade's would have been better still if they were corrected for this bias.[14]

What's clear from Good Judgment's research so far is that there is no single trick to superior forecasting. Rather it's a combination of things – including training, putting forecasters in teams, aggregation and weighting methods – that each increase accuracy by 10% or so. But these add up – put it all together and you get impressively reliable predictions.

§3

Accuracy or usefulness: pick one

Prediction polls and markets disallow pundits' verbiage by asking direct yes-no questions on highly specific events. But there's a 'rigour-relevance' trade-off: unambiguous are not always the most useful.

"You want questions that are rigorous enough within a tournament environment so that they can be scored irreducibly," explains Regina Joseph, who also helps write questions for the Good Judgment Project. "But you have to balance the rigour of the question versus the actual policy relevance question. The central function within the question-formulation team is trying to understand to balance rigour and relevance."

Asking questions that are unambiguous – that will be resolved as correct or incorrect – is not as straightforward as it may seem. "Major markets, to no fault of their own, regularly have had to claw back questions in which some kind of unplanned contingency occurred," says David Rothschild. Even events as apparently binary as election results have been called, only to be reversed soon after – with markets sometimes paying out on the original calls, and in other cases reversing their payments.

Robin Hanson meanwhile suggests that Good Judgment's questions are more 'interesting' than they are useful. Hanson suggests that questions, in order to be useful, should be much more intimately involved with decision-making: less a forecast on what may happen, and more an immediate input to actual decisions. "Useful questions are close to a decision, they're actionable," he says. "The farther they are from a choice and its consequences, the less useful a prediction is."

Hanson suggests more decision-relevant questions – such as the consequences of taking a certain action – should be asked. Good Judgment is, in its final stages, now experimenting with such conditional questions: for example asking whether or an event will occur, given different actions taken by government.

Good Judgment is limited in its questioning scope by not being able to ask questions directly relating to US affairs, either events transpiring there or regarding its policies. This is the result of a fiasco in the early 2000s, when DARPA first experimented with prediction markets. The agency commissioned a group – involving Hanson – to run a market forecasting Middle East events after 9/11, what was soon labelled as 'terror futures', with participants potentially profiting from negative outcomes ('betting on death', though Hanson reiterates that wasn't the case). The head of DARPA head resigned and the project was killed. "DARPA felt like their hands were slapped there," says Hanson. "That's an unusual degree of negative publicity."

As a result, the value of the questions asked at Good Judgment is limited, says Hanson. "The most important thing about a question is that it be something you care about the answer to," says Hanson. "Even accuracy is less relevant than how close to a decision it is."

Joseph claims however that question designers have got better at achieving this rigour-relevance balance – and "cracked that nut". Every year there has been a consistent improvement in the policy relevance of question content at Good Judgment, she says.

In order to design a useful polling or market question, you should start with what you need to know – will this technology be a success, will there be turbulence in a certain part of the world? – and then work back to specific, quantifiable events that you know will be verifiable. This is limiting to an extent, to a restricted subset of the world of 'known unknowns'. "It's easier to write questions that are simple and have outcomes that are verifiable by data that you already have," says Rothschild. "Which are not necessarily the most interesting and executable things."

Then again, he says, if it is hard to write a polling or market question about a certain event, then it will also be that much harder to write a statistical model that can solve it computationally or any other way. It is a delicate balance. Human brains are often much better at making sense of complex, ambiguous futures than data-driven approaches (which excel in well-defined, regular universes where the future looks much like the past). Precision is not the only valuable measure. When it comes to navigating a world of black swans and unknown unknowns, 'precise' tools shouldn't

fully displace long-term scenario planning that embraces complexity and employs narrative and imaginative techniques.

§4

Discovering your inner superforecaster

The headline result from Good Judgment has been that many individuals really are good at forecasting, and reliably so: forecasting is a skill, and a skill that can be developed and honed. The better the individual forecasters, the better the group judgment – but what can we learn about better subjective forecasting in general?

"Some forecasters are just better than others," says Pavel Atanasov. "And that seems to hold from question to question, and year to year. So we've been spending some resources just figures out what makes for those better forecasters."

Good Judgment found that the best forecasters scored higher on tests of cognitive abilities (fluid and crystallised intelligence); had more open-minded cognitive styles of thinking; had training in probabilistic reasoning and in overcoming common biases; worked in collaborative teams rather than alone; spent more time thinking about questions before making forecasts; and revised their forecasts regularly:

- Fluid and crystallised intelligence

Fluid intelligence refers to cognitive processing, individuals' ability to detect underlying patterns and to "adapt readily to new problems and find good solutions".[15] In forecasting, this amounts to the ability to draw parallels between a current situation and the evidence in front of you with a historical analogy, spotting similarities and differences. This also encompasses the capacity to resist instinct and engage in second-order, more considered and rational 'system-two' thinking.

Yet you cannot draw such parallels if you know little about the relevant subject matter. That kind of intelligence is 'crystallised intelligence', and refers to actual knowledge and the ability to use it. To measure it, the Good Judgment researchers outright tested people on geopolitical knowledge, asking questions such as: 'Does the leadership of Saudi Arabia, like that of Iran, embrace the Shiite branch of Islam?', or 'Does the World Trade

Organisation view agricultural subsidies as a major problem in moving forward with trade liberalisation?'.[16]

High marks on tests of these two types of intelligence was found to be a good predictor of reliable forecasters. Both, working together, are important, says Atanasov. "You need to have background knowledge in order to know where to start your search [crystallised intelligence], but also need to know when to connect the dots when you gather evidence, discerning true patterns from spurious ones [fluid intelligence]."

- Open cognitive style

When considering how people prefer to and tend to think naturally, researchers have suggested that the best forecasters are associated with cognitive styles of thinking that 'tap into openness'.[17]

The Good Judgment researchers measured various ways of thinking that they believed would predict forecasting accuracy, including individuals' openness to experience and the way they handled ambiguity. They found that more 'actively open-minded' forecasters performed better: good forecasters actively search for reasons why they might be wrong, rather than ignoring or discounting contrary evidence, and update their views in light of new information.

Much of this may be common sense, but whilst previous experimental evidence has linked actively open-minded thinking to estimation accuracy, Good Judgment's experiments are the first to demonstrate this with the forecasting of real-world problems.

- Training

One of the most encouraging findings of Good Judgment's research so far is that a very small amount of training goes a long way to improving forecasting ability. "When you give people training the average person is 10-15% more accurate over the long run," says Atanasov. That's impressive given that it's just an hour's training: some participants go on to spend more than 100 hours forecasting a year.

Training is straightforward, too. It involves explaining what biases, heuristics and psychological temptations may sway judgments off the mark rather than help, and how to spot them, as well as training in probabilistic reasoning: on grounding estimates in relevant base rates, and how to 'be a good Bayesian', updating beliefs in response to new information.

- Teaming

The final factor that greatly improved individuals' forecasting accuracy was their assignment and involvement in forecasting teams. Whilst there are many distorting forces that can cause judgment in groups to err ('process losses': including pressures to conform, overweighting common information, poor coordination and conflict), there are also many benefits ('process gains': sharing information, individuals' spotting others' biases, motivation, and the gain from diverse opinion). Researchers who have focused on dedicated forecasting environments such as Good Judgment have found that the gains exceed the losses. Lone forecasters are less accurate.

Identifying the best forecasters is a combination of looking at their predictive record so far (past success being a reliable, though not flawless, indicator of future success) and at these factors above.

But perhaps the most significant factor in individual forecasting success – and so in turn, the success of these collective forecasting tools – is dedicated practice. By answering hundreds of questions on a regular basis, and being scored on each, has given participants black-and-white feedback on their forecasting capabilities. Through such feedback they can see for themselves whether they tend to be overconfident about certain events, or too quick to make decisions, or too susceptible to certain kinds of evidence – and from this, they can then try new strategies and make better forecasts. With each forecast, they learn to be better with probabilistic reasoning.

This deep deliberative practice is essential for developing forecasting expertise, say researchers. But its cultivation at the Good Judgment Project is unique, and it is hard

for organisations to simply set up their own prediction tournaments on one-shot forecasts and expect the same kind of predictive power. Can these tools be realistically adopted in the real world?

§5

Harnessing collective wisdom in real-world decision making

As much as it is a research operation, Good Judgment is not an entirely artificial environment: people are joining from all over the world, answering questions on real-world events of real-world importance. It is simply one dedicated to forecasting, which organisations could employ – tapping into a network of suitably incentivised and involved forecasters. (Good Judgment itself will be spun out in such a way after the research period ends.)

However, organisations can't easily set up their own forecasting tournaments as well-funded and well-resourced as the Good Judgment Project. "Several aspects of Good Judgment's rigorously scientific composition will make it extremely difficult, expensive and time-consuming to match," suggests Regina Joseph in a paper on the tournament's design.[18] Nonetheless, says Pavel Atanasov, Good Judgment has invested in exploring new methods and running experiments, the results of which others can simply adopt more or less off-the-shelf. Many of the tools and techniques that have been developed can be employed at low cost, and apply equally well to small-scale tournaments, he suggests.

Robin Hanson contends that, whilst "they're certainly learning useful things" at Good Judgment, scaling some of its polling approaches may prove difficult (and in these cases, markets will fare better). As discussed in §1, Good Judgment has concentrated on making raw forecasts as good as possible, as well as exploring various ways of tweaking its polling algorithms to correct for biases. "That's just going to be hard to scale to a larger, more flexible world," says Hanson. "Once you imagine much larger systems and people disagree about what biases are at work, then I think the financial market approach makes more sense." The promise of markets in this case would be that there is always an incentive for traders to look for biases and correct the market, profiting in the process.

Use in corporates

Prediction markets and polls have been experimented with in companies, and used by decision-makers, for more than a decade. Their use is more directed towards

information gathering, as a tool for open innovation and crowdsourcing, than towards strategic forecasting.

Several case studies have shown how prediction markets can be very powerful in business.[19] In short, they're useful for aggregating what employees know, and particularly what employees might know but not reveal through normal channels – for fear of retribution, and so on. They're particularly good at predicting things such as project-completion dates, says Robin Hanson: where the employees 'in the trenches' know the relevant information. "They are especially useful in situations where, if you just ask people, people won't tell you the truth."

Despite much hype, mainstream adoption has been poor. Whilst many companies have run them, they have not been used as much to inform high-level decision-making. Hanson suggests that many companies have set up markets poorly, asking interesting but not useful questions – judging markets by activity and 'buzz', but not accuracy or actionability.

Beyond being able to use these methods in day-to-day decision-making, and reallocate resources accordingly, there is, says David Rothschild, a question of whether or not decision-makers are comfortable with highly quantifiable, fast and accurate predictions: as it might disrupt the job they're hired to do.

Hanson describes the experience of the US' Missile Defense Agency, which runs missile tests. "Often they schedule a test and then it doesn't happen, because one of the parts – missile, launcher, people ready to watch where it goes – wasn't ready," he says. Such tests costs millions of dollars to set up, and millions are lost when tests are scrapped. Knowing, further ahead of time, whether a test was going to go ahead or not could save a great deal of time and resources. For that reason, the agency set up a prediction market to forecasts whether tests would go ahead on schedule. The predictions worked. "But that embarrassed people and they got the project killed, just like with other project-deadline forecasting in other industries," says Hanson. "People looked bad because they didn't have the excuse that 'no-one could have foreseen that'."

As an article published by Deloitte on prediction markets for corporate foresight notes:[20]

> The technology is easy and widely available. The design of a program with incentives, recruitment of participants, good forecasting questions, and alignment to a company's culture may not be easy to achieve. While many prediction market vendors are probably too ready to sell you their market and technology as a solution, what is really important is their capacity to support the organizational acceptance of the technology.

In companies such as Microsoft, there are signs however that change is afoot, says Rothschild, and employees at all levels are learning how to fit such methods into their workflow, budgets and decision-making. "But it's moving much slower than you'd probably think from the outside."

Conclusions

Prediction polls and markets are forecasting tools that are useful when you want to elicit and aggregate information and expertise that is dispersed amongst a group. For forecasting, if you can aggregate individuals' perspectives the right way, each brings different bits of actual knowledge of the antecedents that can add up to produce a more complete picture of the future. In markets, a profit incentive promises that people will always seek out the best available information.

In the face of an uncertain and complex future, these methods will be useful when information or expertise that can help make a prediction is 'out there' in the world, and distributed among many people. They don't, as such, apply to all kinds of question. (As Robin Hanson puts it, "If you want to know what I had for breakfast, just ask me.") There are many cases where the social phenomena in question are more regular, 'clocklike', and so mechanically predictable – in those cases, sometimes statistical models can crunch the data better than human judgment. You don't, for example, need to establish a prediction market to predict a football match: the available data, statistical modeling techniques and computing power we have now means that predicting such events to a high degree of accuracy is increasingly trivial.

Data-driven approaches are powerful when the data is well-structured. When it comes to predicting the future, that in practice requires that you have a well-defined sampling universe – some regularity such that the future will look much like the past. Faced with complex, ambiguous futures, human brains are often much better at making sense of the information available to them. And the best way of benefiting from human judgment is to use the judgments of many.

Organisations who familiarise themselves with these tools, and adopt them into their ways of working, will reap many benefits. These are tools that are valuable not just for forecasting, but for eliciting information that is dispersed amongst an organisation that is not easy to get directly or in other ways. They're useful when you have reason to suspect that the usual channels are failing, when people aren't telling you what they really think, or when you aren't sure of whom the individuals are who know the answer to the question.

But research into these tools, and the underlying principles, will have wider influence: improving how we can better elicit and aggregate information, crowd-source ideas, and improve polling and market research.

Market research, in particular, has much to learn. “Survey research is a multi-billion dollar fraud,” says David Rothschild. Most research firms run surveys with the same kinds of questions that were asked decades ago, with selection and aggregation methods that are “ancient and extremely unrobust,” he says. One area Rothschild is particularly interested in is ‘expectation polling’, in which, as opposed to opinion polls, people are asked what they think will happen. The evidence suggests that this approach can be much more reliable than opinion polling.[21]

*

There’s more research to be done to improve human collective judgment more still. As Good Judgment nears its close, it’s running some more speculative experiments: for example placing its superforecasters in prediction markets (‘supermarkets’). The idea of markets full of highly-engaged superforecasters is interesting because markets are normally thought to need smart marginal traders, ‘sharks’, as well as ‘fish’, who place a bet here and there but lose money – feeding the sharks. What happens in a market in which everyone is a shark?

Good Judgment has also recently started experimenting with ways of making “fuzzier” predictions and conditional forecasts (if this happens, then will that happen?). If these forecasts turn out to be reliable and accurate, they’ll be much more useful.

But prediction polls and markets have amply demonstrated to be powerful predictive tools. Robin Hanson suggests that rather than funding yet more research, organisations should jump in and adopt these ideas more seriously. “We’ve had the ability for a while,” he says. “There’s a lot of interest in funding research, but not so much interest in actually doing them.”

Digital platforms make it easy to set up polls and markets, internally or for the public. And indeed, many services exist allowing for companies to use public markets, and offer some expertise (with appropriate fees) for setting them up with some rigour. Software for organisations to set up their own tools, such as Augur, an open-source platform for prediction markets, is readily available. But it is clear that it is not the cost or availability of such infrastructure that poses a barrier to their adoption. Asking the right questions, that will not only resolve properly but be useful and actionable to an organisation, takes expertise – and there is a significant marginal cost in setting up additional questions.

Human factors

The importance of the individual forecasters' capabilities, for better crowd forecasting, is enormous, says Pavel Atanasov. "Better foresight from these tools depends on the people," agrees Regina Joseph. "They don't generate good foresight in and of themselves."

One trend in the research is particularly clear: the more practice individuals have at forecasting, the better they get. In the highly-engaged, dedicated tournament environment of Good Judgment, forecasters get regular, numerical feedback on their skills – they see what they got wrong and right, and do better next time. This kind of practice and training is absent in everyday forecasting – of the kind businesses and individuals make, implicitly and explicitly, in every action they take. The benefits of better judgment are obvious, but the opportunity to realise it is perhaps greater than has been previously thought.

"I believe improving the human factors created an important advantage for the Good Judgment Project," concludes Atanasov:

> After potential forecasters pass our screenings, we invest in training them and helping them cultivate their skill. For the most part, they respond by putting in the requisite effort in hunting down information, updating their predictions and engaging in discussions with peers. Put shortly, forecasters cared.

Building a sophisticated, state-of-the-art crowd prediction system is impressive. “But unless you recruit, motivate and train human forecasters to use it, it would not produce very accurate forecasts.”

References

Arrow, Kenneth J., et al. "The Promise of Prediction Markets." *Science* 320 (2008): 877–878.

Atanasov, Pavel, Philip Rescober, Eric Stone, Samuel A Swift, Emile Servan-Schreiber, Philip Tetlock, Lyle Ungar, Barbara Mellers. "Distilling the Wisdom of Crowds: Prediction Markets versus Prediction Polls," (2014), draft.

Bikhchandani, Sushil and Sunhil Sharma. "Herd Behavior in Financial Markets: A Review." *IMF Working Paper No. 00/48* (2000). Retrieved at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=228343

Cowgill, Bo and Eric Zitzewitz. "Corporate Prediction Markets: Evidence from Google, Ford, and Firm X." 2014, draft.

Hayek, Friedrich. "The Use of Knowledge in Society." *The American Economic Review* 35, no. 4 (1945): 519–530.

Joseph, Regina. "Keeping Score: Better Policy Through Improved Predictive Analysis." *5th International Conference on Future-Oriented Technology Analysis* (2014).

Kambil, Ajit. "Social analytics: Tapping prediction markets for foresight." 2010. Retrieved at: http://www.lumenogic.com/www/static/pdf/deloitte-kambil.pdf

Landemore, Hélène and Jon Elster, eds. *Collective Wisdom: Principles and Mechanisms*. Cambridge University Press, 2012.

Mellers, et al. "The Psychology of Intelligence Analysis: Drivers of Prediction Accuracy in World Politics." *Journal of Experimental Psychology: Applied*, forthcoming.

Page, Scott. *The Difference*. Princeton University Press, 2008.

Rothschild, David. "Comparing Prediction Markets, Polls, and their Biases." *Public Opinion Quarterly*, 73, 5 (2009): 895–916.

Rothschild, David and Justin Wolfers. "Forecasting Elections: Voter Intentions versus Expectations." Brookings, 2012. Retrieved at: http://www.brookings.edu/research/papers/2012/11/01-voter-expectations-wolfers

Servan-Schrieber, Emile. “Prediction Markets: Trading Uncertainty for Collective Wisdom,” in Landemore and Elster (2012), 21–37.

Sunstein, Cass. “Deliberating Groups versus Prediction Markets (or Hayek’s Challenge to Habermas).” *Episteme* 3, 3 (2006): 192–221.

Surowiecki, James. *The Wisdom of Crowds*. New York: Anchor, 2005.

Tetlock, Philip. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press, 2005.

[1] Tetlock (2005).
[2] See in particular Surowiecki (2005) and Landemore & Elster (2012).
[3] Hypermind's own figures and selection of forecasts. Available at: http://blog.hypermind.com/2014/11/15/2014-us-midterm-elections/
[4] Atanasov et al. (2014), p9; Joseph (2014), p6.
[5] Hayek (1945). Note that Hayek would be unlikely to endorse markets simply as a means of knowledge aggregation that can yield a single 'right' answer.
[6] A 2008 *Science* article, 'The promise of prediction markets', features a who's-who of economics arguing for looser betting regulation to allow prediction markets – Arrow et al. (2008).
[7] See e.g. Surowiecki (2005) and Sunstein (2006).
[8] Sunstein (2006), p200-202.
[9] See Bikhchandani & Sharma (2000) for a review.
[10] Servan-Schrieber (2012), p32.
[11] Page (2008).
[12] Union Square Ventures VC Fred Wilson acknowledged this with his 2015 predictions: http://avc.com/2015/01/what-is-going-to-happen/
[13] Atanasov et al. (2014), p22.
[14] Rothschild (2009).
[15] Mellers et al. (forthcoming), p6.
[16] Ibid, p7.
[17] Ibid, pp10-11.
[18] Joseph (2014), p4.
[19] Cowgill and Zitzewitz (2004).
[20] Kambil (2010).
[21] Rothschild (2012).