THE WILBERFORCE SOCIETY

Prediction Markets and Forecasting Techniques

Editor: Dan Moinard

Writers: Matija Franklin, Kartik Vira, Khuzaimah Saeed, and Nikolas Bernaola Álvarez

The Wilberforce Society | www.thewilberforcesociety.co.uk

Cambridge, UK | April 2020

EXECUTIVE SUMMARY

This paper explains that public institutions, businesses and individuals can improve the accuracy of their forecasts, and suggests how to do so. It begins by summarizing the current state-of-the-art research on subjective judgement in the context of prediction, with applications ranging from geopolitical events to economic trends. Recent studies have produced two main results regarding how individuals and groups can improve the methods used when making forecasts. First, they explain how to consider people's cognitive biases in forecasting. Second, they show how to create context-appropriate techniques to elicit information and refine their judgement.

This paper analyzes several key areas in detail:

Evaluating forecasting: Starting with the requirements for good and verifiable forecasting, methods for assessing and scoring forecasts are introduced. Tools to monitor forecasts can be used to extract cognitive and methodological characteristics of forecasters, such as under-confidence or over-confidence, and in particular those of the most accurate forecasters.

Cognitive biases: Next, with a first-principles approach, the psychological processes at play in subjective judgement are presented. A range of individual cognitive biases can explain the strengths and weaknesses of subjective judgment; some additional biases apply specifically to groups. In both cases, precautions can be used to directly counteract common heuristics. These schemes impact general information aggregation methods and their implementations.

Prediction markets: All of these concepts are then applied to study prediction markets, stock market analogues for trading contracts on the outcome of specific events, which are a promising solution for information aggregation. Prediction markets tend to be efficient under a wide range of conditions, in part because they encourage participants to avoid cognitive biases. Their recent implementations have shown success in predictions ranging from political election results to corporate performance.

Long-term forecasting: The paper then examines an important current limitation for forecasting, long-term predictions, and explores the specific issues that appear in this context. While defining "long-term" requires context-dependent timescales, some tools are available to deal with general difficulties, for instance the huge number of possible scenarios leading to an

event. One must also take into account so-called "black swans", very-high-impact but low-probability events that may fall outside the scope of forecasts. Predictions of catastrophic events and existential risks, including those related to artificial intelligence, are discussed as the most extreme of forecasting objectives.

The paper concludes by discussing how these various areas relate to the future and progress of forecasting. A wide range of policies are suggested, aimed towards the increased use and improvement of forecasting, applicable by governmental agencies, private companies and individuals. The proposed policies and techniques have the following goals:

1) Incentivize and raise awareness for appropriate forecasting methodology, from basic training to funding pivotal studies in large-scale organizations;

2) Create tools and frameworks that allow people to avoid their individual and collective biases, such as heuristics checklists and validation tools;

3) Implement systematic checks and feedback for all predictions made within organizations, including governmental forecasts and private companies' financial guidances;

4) Subsidize research and high-impact real-world applications of various forecasting techniques, and in particular prediction markets, by strengthening their legislative framework and creating plug-and-play prediction systems for private companies;

5) Encourage the public release of prediction data and forecasts from private companies, promoting a culture of progress and transparency.

Specific policy recommendations are suggested to reach these goals. Finally, the paper's conclusion places the topic of forecasting techniques within the larger area of decision-making, and how improved predictions may contribute to progress.

TABLE OF CONTENTS

I. INTRODUCTION

I.I. INTRODUCTION TO FORECASTING AND EXPERT JUDGEMENT

> *"People who spend their time, and earn their living, studying a particular topic produce poorer predictions than dart-throwing monkeys who would have distributed their choices evenly over the options."* – Daniel Kahneman

In a now-famous study, Philip E. Tetlock found that even the most intelligent analysts were as accurate as "dart-throwing chimps"[1]. This inspired him to co-create the Good Judgment Project (GJP), a large scale study of forecasting accuracy. The project was part of the Aggregative Contingent Estimation (ACE) program run by the Intelligence Advanced Research Projects Activity (IARPA), which funds research for the US intelligence community. The goal of this project was to approach forecasting from a scientific perspective, by taking into account people's psychological shortcomings, making predictions more precise in terms of wording and measuring the success of these predictions more systematically. Tetlock described the success of the GJP in his book Superforecasting[2]. The GJP's legacy is that large-scale studies can be used successfully to yield concrete results which improve forecasting ability, measurement and requirements.

These results, which will be discussed at much greater length in this paper, lead to a much broader realization: every person uses forecasting and subjective judgement skills every day. In personal and professional settings, we all make predictions daily, both implicit and explicit, such that forecasting abilities can be considered as tools for everyone, rather than simple curiosities for experts. In this context, the GJP results demonstrate that we, as individuals and as a society, are quite bad at forecasting, but that we know how to assess these skills properly. Most importantly, we know how to improve.

Making accurate and reliable predictions is of course of great personal value to each individual. Any choice that we make relies, at least implicitly, on a subjective judgement concerning the outcome of an action or decision. This applies to contexts as varied as relationships, financial planning and safety concerns. For instance, when leaving your home, you will systematically ask

---

[1] Philip E Tetlock, *Expert Political Judgment* (STU-Stud, Princeton University Press 2005).

[2] Philip E Tetlock and Dan Gardner, *Superforecasting: The Art and Science of Prediction.* (Crown Publishers/Random House 2015).

yourself whether you need an umbrella. This forces you to make an implicit prediction about the chance that it will rain while you are outdoors. You will then collect information, by checking for menacing clouds in the sky, asking the advice of your partner, checking some weather website, etc. Given that information, you will then make a judgement based on your view about these sources of information. Forecasting skills are therefore constantly being used, at least at a subconscious level.

Beyond everyday contexts, accurate forecasting has an important societal value. Predictions are made and relied on in all professions. These include weather forecasters, policymakers, military strategists, news pundits and so on. Such a significant part of their daily activities rely directly on forecasts in some form. One should also note that the most explicit forecasts are typically made for the highest impact events, such as the likelihood of a country going to war, the probability of a global catastrophe, a country's projected GDP growth, or the expected gains from a new technology.

This huge range of applications also implies that, up to now, an immense majority of decisions have relied on forecasts whose accuracy and basic principles had not generally been thoroughly studied. It is therefore reasonable to expect that treating forecasting with a more scientific and precise approach will lead to more informed decision-making, based on less biased and noisy forecasts and a better understanding of their results. This would in turn create a huge potential for adopting better strategies for mitigating risks and enhancing positive effects.

I.II. THEORETICAL INTRODUCTION: DEFINITIONS

What is meant by "forecasting techniques"? In this paper, they are defined as methods of eliciting and aggregating information in order to form an informed judgement about future events. In other words, all the different ways in which information can be acquired from various sources, and how to judge these sources individually and collectively in order to obtain an insight and quantifiable prediction about an event. Information elicitation and aggregation can be contrasted with data generation and collection, on one hand, and decision-making, on the other. Furthermore, forecasting techniques will only be discussed based on subjective judgements, as opposed to pure modelling, which attempts to find mathematical models suitable for deterministic or statistical predictions.

Both the weakness and the genius of forecasting lie in the subjective nature of the judgements required to obtain a prediction. An immediate consequence is that, in order for forecasting techniques to be useful, they need to yield verifiable predictions. Indeed, since the goal of predictions is to provide judgements and insights about some future occurrence, it is crucial to be able to check the outcome of the forecasted event.

It is therefore possible to list the basic requirements of well-defined forecasts. They need to 1) predict the outcome of some verifiable event, 2) give a specific deadline, and 3) establish a clear resolution criterion or an entity responsible for determining the outcome. For instance, weather forecasts such as "there is a 70% chance of rain in Cambridge tomorrow" fulfill these conditions. An example of an ill-defined geopolitical forecast could be "there is a 60% probability that the war in Iraq will end", as it lacks a specified deadline.

An important step forward is the application of a full set of scientific methods to forecasting, as was started by the IARPA studies. This is analogous to the introduction of randomized trials in medicine. For instance, such studies could help us understand under which conditions opinion polls are reliable or how best to aggregate the opinions of experts on a specific subject. Novel methods based on Big Data have shown promise, such as bibliometrics that analyze academic publication and patent approvals to predict the growth of a given technological area. Testing their accuracy and limits in a scientific way would allow them to be improved, established as reliable tools and used widely[3]. Other more established methods such as Delphi iterative surveys, used for industry-wide or national forecasts, may also be improved. A typical limit of Delphi approaches is the difficulty of assigning different degrees of confidence to the experts interviewed based on a subjective assessment of their level of expertise. Obtaining a proper calibration through some other means, for instance their participation in a prediction market, would significantly reduce the assumptions and uncertainties in such research[4]. In particular, Tetlock discussed the impact of selecting well-calibrated experts in polls to avoid overconfidence[5].

[3] Ayse Kaya Firat, Stuart Madnick and Wei Lee Woon, 'Technological Forecasting–A Review' (2008).
[4] Alan L Porter and others, 'Technology Futures Analysis: Toward Integration of the Field and New Methods' (2004) 71 Technological Forecasting and Social Change 287.
[5] Ville A Satopää and others, 'Probability Aggregation in Time-Series: Dynamic Hierarchical Modeling of Sparse Expert Beliefs' (2014) 8 The Annals of Applied Statistics 1256.

I.III. OBJECTIVE OF THE PAPER

In this paper, it is argued that the forecasting techniques defined above can and should be improved. The primary aim of this paper is therefore to raise awareness for this subject and to summarise the ways in which improvements can be implemented. The heightened awareness of the topic could encourage individuals and institutions to review their attitudes towards these issues, leading to the promotion of good forecasting practices. Most importantly, these would include avoiding vague phrasing in predictions and dispelling public misperceptions.

This involves providing a summary of current forecasting techniques and the subtleties put forward by recent academic studies. Among these, the understanding of the psychological processes at play when making subjective judgement and decision is essential. Not only is this topic the most widely applicable to individuals and groups, but these cognitive processes are the fundamental building blocks of all subjective judgements in forecasting, and therefore of any potential improvement. Our secondary objective is to participate in the development of a framework for accurate forecasting. The implicit goal here is to help the transition of forecasting from a purely subjective and artisanal craft to a systematic and scientifically-backed area of study and work.

With important research results starting to be released, it is important to aggregate these findings and use them to create some kind of generalized framework, through which more precise and subject-specific studies and policies can be examined. This article has therefore been designed as a policy paper. Beyond the summary of the current state of forecasting, the output of this paper is a set of broad recommendations and specific policy proposals. These could be implemented to produce better predictions in corporate, academic and governmental settings.

I.IV. OUTLINE OF THE PAPER

This paper starts by defining the requirements for good forecasting and verifiability, introduces methods for assessing and scoring forecasts, and uses these to discuss in more detail the results obtained by the GJP. Next, it presents the psychological processes at play in subjective judgement, in order to create a first-principles approach. Starting from individual cognitive biases, it discusses how psychological concepts can explain the strengths and weaknesses of subjective judgment. Their application to groups rather than individuals is then reviewed, concluding on their impact on general information aggregation methods and their implementations.

All of these concepts are then applied to study prediction markets, stock market analogues for trading contracts on the outcome of specific events, which are a promising solution for information aggregation. In particular, this paper discusses why prediction markets tend to be efficient, how they avoid cognitive biases, and how they have been implemented in recent years for predictions ranging from political election results to corporate performance. While the choice to analyze prediction markets is deliberate, it should not be understood as an exclusive endorsement of that method over other forecasting techniques: rather, the proposed analysis of prediction markets in the light of its relation to psychological biases could analogously be made for a wide variety of other forecasting techniques.

This leads to an examination of an important current boundary for forecasting: long-term predictions. The first topic of discussion focuses on the importance of timescales and what is meant by "long-term", explores what specific issues appear in this context, and what possible solutions are being explored. An additional consideration are so-called "black swans", highly improbable events that may fall outside the scope of forecasts but whose potential impact make them too important to ignore. All of these concepts and more are illustrated through the example of forecasts of catastrophic events and existential risks, reviewing their intrinsic difficulties and considering cases such as the advent of artificial intelligence.

Having reviewed the main current issues in forecasting, policy recommendations are formulated. Some wide-ranging policy goals are argued, setting a path for the rapid improvement of forecasting techniques. Then, a number of specific policy proposals are made to fulfill each of these goals. These proposals span a large number of contexts, including individual and professional environments, and public and private institutions.

Finally, the conclusion of this paper discusses the difference between forecasting and decision-making, reviews the main conclusions of the paper, and reiterates the primary goals towards improvement of prediction methods. This policy paper closes by exploring the importance of such an improvement, its effect on the current state of society and its impact on future generations.

II. EVALUATING FORECASTING

II.I. FORECASTING METHODOLOGY

In order to use forecasting techniques that can be judged, verifiable predictions must be possible. To reiterate, well-defined forecasts need to predict the (probabilities of each) outcome of an event, within a specific deadline, and have clear resolution criteria or an entity responsible for determining the outcome.

As shown in the examples discussed in the introduction, there exists a variety of different prediction types. In particular, these can be separated into two distinct categories, based on the category of outcome:

- Binary forecasts will assign a probabilistic value to the "True"/"False" outcome of an event; the prediction will then answer the generic question "What is the probability of event A happening?"
- Continuous-outcome forecasts assign a numerical value to some variable; these include the date at which an event is predicted to occur, or what the value of a certain metric will be at a given time.

The two concepts are, however, intimately related, as a continuous-outcome prediction can be re-written as an infinite set of binary-outcome predictions. However, this mathematical relationship is only partially applicable in practice, an important limitation we will discuss.

The majority of the existing literature deals with binary forecasts, as these are the simplest conceptually and offer the easiest verification through a standard "yes"/"no" answer. However, there are multiple ways in which simple binary predictions can be used to refine the forecaster's view. For instance, as is currently used in sports betting, binary forecasts can be merged to form aggregate forecasts: the payoff will only occur if all, or at least a predetermined number, of the binary predictions are realized. It may still seem that complex forecasts are prohibited under this scheme: the future value of a stock price, the expected temperature on a given day, the projected sales for the coming financial year, the casualty estimates due to a catastrophic event, among others, cannot be expressed directly as a "yes"/"no" answer.

One solution is to split the possible values for these continuous variables into multiple sections. For instance, instead of predicting that the UK GDP growth will be equal to 2% in 2017, a forecaster could express two overlapping binary predictions:

- A 20% chance that growth will be below 1.8%
- An 80% chance that growth will be below 2.2%

These two predictions actually provide more information than simply stating "growth will be 2%", as it creates a rough probability distribution. By eliciting a larger number of such binary predictions from forecasters, their implicit probability distributions can be sampled. Such a system also implies simplicity in assessment: each prediction can be scored individually, as we will now describe.

II.II. ASSESSMENT AND SCORING PREDICTIONS

Having detailed the importance of forecast verifiability and described how to ensure that predictions are well-defined, one can now turn to assessing and scoring predictions. The main goal here is to enable the study of different forecasting techniques or forecasters. Indeed, to obtain the necessary evidence to conduct such studies, we need to answer the question: how accurate is a forecast? In order to answer this question, the introduction of metrics is necessary for the assessment and scoring of both single predictions and sets of predictions.

The main purpose here is to standardize the assessment mechanism: this is a clear requirement for any objective comparison of forecasts, forecasters, methods, conditions, and so on. More generally, the ability to score forecasts is a prerequisite to any scientific approach on the subject. It should be noted, however, that this process only evaluates the accuracy of a forecast and ignores the value of the particular forecast being accurate.

The most widely-used metric for scoring the outcome of a prediction is the Brier score. It applies only to binary and multi-category events, generally takes a value between 0 and 1, and increases as the square of the error:

- A perfect prediction, forecasting 100% probability for the event that occurs, yields a Brier score equal to 0;
- A completely incorrect prediction, forecasting 0% probability for the event that occurs, yields a Brier score equal to 1;

- A neutral 50%-50% forecast gives a Brier score of 0.25, irrespective of outcome;
- If a forecaster states a 70% chance of rain for today, and it does rain, their Brier score is $(30\%)^2$=0.09.

Two important extensions to this single-question, single-timeshot scoring method are possible. The first is to compute the Brier score of a set of forecasts (or alternatively, of a group of forecasters for a single event) by averaging the Brier scores over forecasts (or forecasters, respectively). The second is to compute the Brier score of a forecast that was updated over time, before the event outcome. This could be the average of updated forecasts over time, or a weighted-average with more recent forecasts being weighed more heavily.

While a single Brier score allows us to assess accuracy for a single forecast, its true value lies in the use of statistical data: there is significant information in the Brier scores for a large set of predictions. The Brier score can be decomposed into 3 contributions:

1. Calibration: how far the predicted probability is from the true probability of the event;
2. Resolution: how close to 0% / 100% forecasts are;
3. Uncertainty: intrinsic uncertainty of events.

The former two are measures of the forecaster's skill, allowing different forecasters or forecasting techniques to be understood, beyond the low-level accuracy given by the Brier score. They can be represented graphically in a scatter plot of predicted probabilities versus true probabilities, i.e. plotting the observed frequency of event outcomes against their predicted probability. Points precisely along the diagonal correspond to good calibration. The graphs below shows typical forecast characteristics of over-confident, under-confident and well-calibrated forecasters.

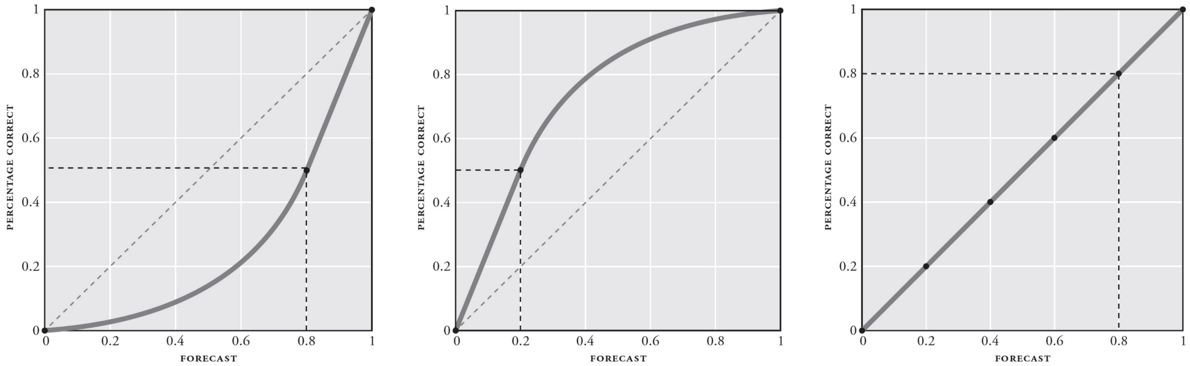

Figure 1: Plots illustrating observed event frequency against forecast probability for an over-confident (left), under-confident (center) and well-calibrated (right) forecaster; reproduced from Tetlock & Gardner (2016)[6].

Having a high proportion of points near 0% or 100% corresponds to good resolution. This means that the predictions are more decisive: the forecaster predicts events that occur with high probabilities, and events that do not occur with low probabilities. The graphs below show the difference between a low-resolution and a high-resolution forecaster.

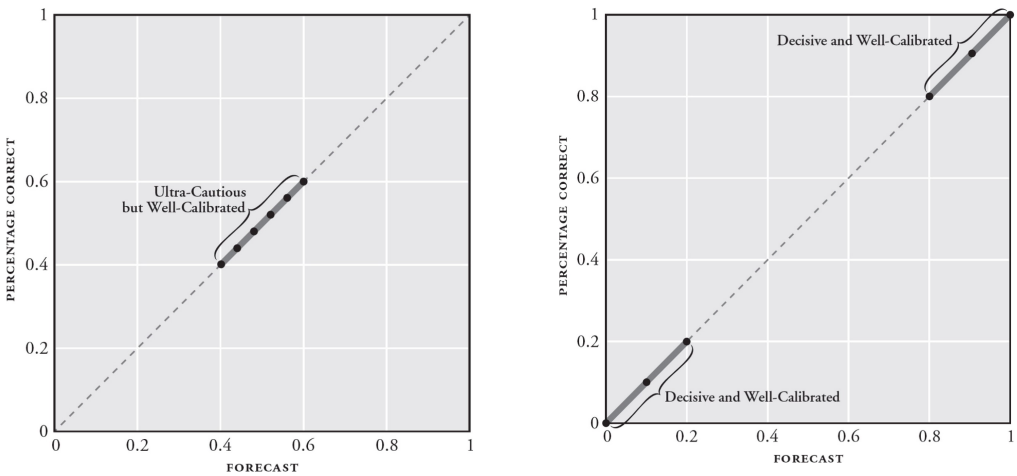

Figure 2: Plots illustrating observed event frequency against forecast probability for a forecaster with perfect calibration, and low-resolution (left) and high-resolution (right); reproduced from Tetlock & Gardner (2016)[7].

[6] Tetlock and Gardner (n 2).
[7] ibid.

Of course, the Brier score has limits, some of which are addressed by alternatives. Most importantly, we note that taking the Brier score of ordinal variables is not possible, by contrast to binary and categorical outcomes. This justifies the phrasing of certain predictions in terms of categories; for instance, we cannot use the Brier score for predicting the value of future GDP growth, but can for any number of ranges of values, such as "below 1.8%", "between 1.8% and 2.2%" and "above 2.2%". Other metrics are also useful, such as the mean absolute error to quantify the distance of the projected quantity (e.g. maximum temperature in the UK today) from the realised quantity (e.g. actual maximum temperature recorded). Different scoring rules also provide other insights: for instance, the logarithmic scoring rule heavily penalizes predictions nearly-100% incorrect[8].

More generally, it should be obvious that there are a variety of metrics that may be more or less useful depending on the context and aims of the studies undertaken. In all cases however, the use of such metrics accomplishes the main goal of standardizing assessment and comparison of forecasts.

II.III. CASE STUDY: EXAMPLE OF GJP FOR GOOD PRACTICES IN FORECASTING ASSESSMENT

The optimal way of illustrating all of these concepts is now to turn to the results obtained during the Aggregative Contingent Estimation (ACE) program, organized by IARPA within the US intelligence community, and more particularly by the Good Judgment Project (GJP) program. This specific example is particularly important as it is widely considered as the seminal work for understanding human judgement in forecasting techniques. More specifically, it is the most comprehensive large-scale assessment of predictions methods used within an organization. While sporadic studies had been previously undertaken to compare the usefulness of certain forecasting techniques, such as foresight in private companies, the ACE program was the first academic study of such scale. The fact that such work was undertaken at the proposal of US intelligence agencies is remarkable. The community is relatively secretive

[8] Tilmann Gneiting and Adrian E Raftery, 'Strictly Proper Scoring Rules, Prediction, and Estimation' (2007) 102 Journal of the American Statistical Association 359.

for obvious reasons linked to the sensitive nature of their work and information, so this project was recognized as a sign of courage and vision.

The main result from the GJP studies is that even forecasting professionals with access to high-quality information make rather inaccurate forecasts: this is the basis for the "dart-throwing monkeys" comparison made by Tetlock[9]. Refining this primary result, it was shown that forecasting accuracy is mostly determined by psychological abilities, rather than subject-specific expertise. In particular, a subset of cognitive skills and traits were determined to be important indicators, such as high open-mindedness and above average fluid intelligence. Another example is use of higher granularity: accuracy was found to decrease when predictions were split into 20% bins for average forecasters (i.e. 0% prediction mapped to 10%, 34% prediction rounded to 30%, etc.), and into 5% bins for the best forecasters[10]. This implies that so-called "superforecasters" analyzed issues with much more nuance and refinement. Similarly, avoidance of certain typical psychological biases was directly linked to accurate forecasts.

These observations linking cognitive abilities to forecasting accuracy have a crucial consequence: they offer ways of improving forecasts. While these will be discussed in detail in the following sections of this paper, important subsets of such improvement methods are debiasing training, exercises to directly or indirectly decrease the effects of cognitive biases on judgement in certain situations, and post-forecast corrections, mathematical operations on forecasts to rectify for known errors and distortions.

The results presented by the GJP also prove that it is possible to successfully assess forecasts by a very large number of individuals and groups alike. Moreover, the singular context of assessing predictions made by the secretive intelligence community indicates that these large-scale studies can be applied at the scope of entire institutions, even those most likely to push back against the effort of transparency that they imply. Perhaps more generally, this body of work showed how research into the assessment of forecasts and the comparison of prediction methods can be achieved successfully and yield valuable insights and paths for future improvement.

[9] Tetlock (n 1).
[10] Tetlock and Gardner (n 2).

III. COGNITIVE BIASES

III.I. HEURISTICS AND BIASES

To make accurate predictions about future outcomes, forecasters need to make a series of decisions about what future outcomes are more or less likely. The cognitive processes that underlie decision-making will, therefore, underlie the cognitive processes needed for forecasting. By definition, decision-making is a cognitive process which results in a selection of a course of action, among various alternatives[11]. Forecasters need to decide what their final prediction is, among other alternative predictions. It is, therefore, important to consider the mechanisms that underlie decision-making, as well as the various factors that might lead to faulty decisions, and therefore inexact predictions.

Heuristics are mechanisms that people use when making judgments. They are mental shortcuts that make individuals focus on one aspect of an issue, while ignoring other factors[12]. Heuristics typically work well, allowing people to make faster decisions. There are certain decision-contexts in which heuristics are useful, such as a baseball player intuitively knowing where to place his hand to catch the ball[13]. As such, heuristics are the underlying mechanism of intuitive, automatic judgments. However, they can also lead to deviations from rational thinking, probability judgments and logic, which are all needed for accurate prediction making. This is why heuristics are problematic for precise forecasting[14], which involves well thought-through decisions. Hence, the automatic heuristic mechanisms could disrupt a forecaster's ability to make slower, more rational decisions. The decision errors that result from heuristics are called "cognitive biases", and many different types of biases exit.

Most cognitive biases are context dependent, and different decision-contexts may trigger different biases. Some biases that can affect one's ability to make a good forecast are the: self-confirmation bias[15], availability bias[16], representativeness bias[17], ambiguity effect[18], attentional

[11] Daniel Kahneman and Amos Tversky, 'Choices, Values, and Frames.' (1984) 39 American Psychologist 341.
[12] A Tversky and D Kahneman, 'Judgment under Uncertainty : Heuristics and Biases Linked References Are Available on JSTOR for This Article : Judgment under Uncertainty : Heuristics and Biases' (1974) 185 Science 1124.
[13] Peter Todd and others, *Simple Heuristics That Make Us Smart* (1999).
[14] Tetlock and Gardner (n 2).
[15] Scott Plous, *The Psychology of Judgment and Decision Making.* (Mcgraw-Hill Book Company 1993).
[16] Amos Tversky and Daniel Kahneman, 'Availability: A Heuristic for Judging Frequency and Probability' (1973) 5 Cognitive Psychology 207.

bias[19], backfire effect[20], base-rate fallacy[21], declinism[22] and gambler's fallacy[23]. To give an example, forecasters can be affected by the availability bias, because information that is more available to them will have a greater effect on their predictions. This is especially problematic because a lot of the information that is available to us is the result of what the mass media is currently reporting. This can skew the predictions of a forecasters towards the information that is most available. Another example is the gambler's fallacy. Here, individuals' predictions future about independent events are altered by the frequency of past independent events. More specifically, if an event has happened more frequently in the past, a forecaster could make the prediction that it will happen less frequently in the future, even if the events are independent of each other. Again, the information which is available to the forecaster about past events could affect their predictions about the probability of a future event occurring, even though those events are not necessarily related.

As a general rule of thumb, forecasters are less likely to make erroneous decisions if they are better informed about the topic. Decisions made before they inform themselves will be less accurate than their later, better informed decisions. This is problematic because the initial decisions can affect the accuracy of the subsequent decisions. In psychology, this is known as the anchoring effect. Anchoring is a specific type of cognitive bias which makes people more likely to heavily rely on the first piece of information offered (the "anchor") when making decisions[24]. In the context of forecasting, this either means that forecasters will rely too much on initial information, or on their initial decisions, when making their final predictions. This can significantly reduce the accuracy of a forecaster's decisions related to forecasting. As a

---

17 Daniel Kahneman and Amos Tversky, 'Subjective Probability: A Judgment of Representativeness' (1972) 3 Cognitive Psychology 430.

18 Deborah Frisch and Jonathan Baron, 'Ambiguity and Rationality' (1988) 1 Journal of Behavioral Decision Making 149.

19 Yair Bar-Haim and others, 'Threat Related Attentional Bias in Anxious and Non Anxious Individuals: A Meta-Analytic Study' (2006) 133 Psychological Bulletin, 133, 1 - 24 (2007).

20 Craig Silverman, 'The Backfire Effect' (*Columbia Journalism Review*, 2011) <https://archives.cjr.org/behind_the_news/the_backfire_effect.php>.

21 Daniel Kahneman and Amos Tversky, 'On the Psychology of Prediction.' (1973) 80 Psychological Review 237.

22 Pete Etchells, 'Declinism: Is the World Actually Getting Worse?' *The Guardian* (16 January 2015) <https://www.theguardian.com/science/head-quarters/2015/jan/16/declinism-is-the-world-actually-getting-worse>.

23 Greg Barron and Stephen Leider, 'The Role of Experience in the Gambler's Fallacy' (2010) 23 Journal of Behavioral Decision Making 117.

24 Adrian Furnham and Hua Chu Boo, 'A Literature Review of the Anchoring Effect' (2011) 40 The Journal of Socio-Economics 35.

result, forecasters should try to make predictions when, and only when, they have done substantial research on the topic of the forecast.

Better read and better educated forecasters will, in general, be somewhat less susceptible to their own cognitive biases. Biases are especially problematic because people are not explicitly aware of their use of heuristics and biases. People's awareness of what is affecting their decisions has been formalised, and can be better understood with the conceptual model proposed by Daniel Kahneman – dual-process model of fast and slow thinking[25].

III.II. SYSTEM 1 AND SYSTEM 2

In his book "Thinking, Fast and Slow", Daniel Kahneman proposes a dichotomy between two modes of thought: "System 1" and "System 2"[26]. More specifically thought can arise in two different ways, or as a result of two different processes. These two processes consist of an unconscious process and a conscious process. System 1 is a fast, automatic and emotionally loaded system of thought that can and does lead to irrational decisions and behaviour. An example of System 1 are the aforementioned biases and heuristics. System 2 is a slow, logical and thought-through system of thought that can lead to rational decisions and behaviour. An example of this system in action is making an argument based on evidence, such as making a forecast. This model of thought is a useful representation of decision-making because it allows for one to think about how thought processes pertaining to System 1 and System 2 can be changed. Explicit System 2 thinking can be changed through persuasion or education, while implicit System 1 thinking can be changed over a longer amount of time through the formation of new habits.

It is therefore evident that a good forecasting intervention would ensure that forecasters rely on ways of thinking that pertain to System 2, rather than System 1. If anything, forecasters should aspire to reduce the influence of System 1 on their thinking. An example of this would be questioning automatic, "gut" reactions. Another example would be taking into account that other people make decisions and behave in ways that are influenced by System 1 thinking. Hence, it is useful for forecasters to take into account both their own biases as well as the biases

---

25 D Kahneman, *Thinking, Fast and Slow* (Farrar, Straus and Giroux 2011).
26 ibid.

of people they are making predictions about. Forecasters cannot predict behaviour based on the assumption that people are behaving as rational agents [27] [28].

III.III. BIASES ARE MORE SALIENT FOR DECISIONS ABOUT THE FUTURE

One aspect of decision-making that is especially problematic for forecasters is that biases have a larger effect on decisions that relate to future events, such as making predictions[29]. Biases become more effective at influencing long-term thinking, thus making decisions more affect-prone[30]. The context of forecasting is psychologically unique because of the increased impact of biases and heuristics on thought[31]. As a rule of thumb, the further away an event is, the less accurate predictions will be. This makes forecasting one of the hardest forms of decision-making because it is less likely that these predictions will be precise. Forecasters should be more sceptical about predictions they made about events in the distant future.

Additionally, thinking about the future is not a thing that comes naturally to most people. A survey conducted by The Institute For The Future found that 53% of people never thought about events that will occur 30 years in the future[32]. Even 27% of people never thought about events that were 5 years in the future. Only 17% of the participants thought about events that were 30 years in the future at least once a week. The survey concluded that a minority of people were highly future-minded. Future research could test whether great forecasters are above-average at being future-minded. If so, this could imply that this is another psychological trait that can, with some amount of accuracy, predict how good of a forecaster someone will be.

III.IV. GROUPTHINK AND GROUP BIAS

One could argue that the aforementioned biases, heuristics and shortcomings of individual decision-making could be alleviated if a group of individuals made decisions together. However, research suggests that these individual shortcomings can be potentially amplified in

---

[27] Amos Tversky and Daniel Kahneman, 'Advances in Prospect Theory: Cumulative Representation of Uncertainty' (1992) 5 Journal of Risk and Uncertainty 297.
[28] Daniel Kahneman and Amos Tversky, 'Prospect Theory: An Analysis of Decision under Risk' (1979) 47 Econometrica 263.
[29] Connie Poon, Derek Koehler and Roger Buehler, 'On the Psychology of Self-Prediction: Consideration of Situational Barriers to Intended Actions' (2014) 9 Judgment and Decision Making 207.
[30] Timothy D Wilson and Daniel T Gilbert, 'Affective Forecasting' (2003).
[31] Tetlock and Gardner (n 2).
[32] Institute For The Future, 'The American Future Gap Survey' (2017).

group settings. As an example, the availability cascade is a self-reinforcing process where a group belief gains more and more perceived plausibility through its constant repetition in public discourse[33]. Such group-level biases are, by definition, more likely to occur when a group of forecasters are attempting to make a prediction together.

The most famous example of decisions going wrong in a group is groupthink[34]. Collective group effort can lead to additional biases that would not and could not occur at an individual level. Groupthink occurs due to people's innate desire for harmony or conformity in a social group[35]. It results in a dysfunctional decision-making outcome. Individuals in groups try to minimise conflict between individuals by reaching a consensus decision. This consensus decision can often be made prematurely. It will often lack in a critical evaluation of alternative viewpoints, which is the very thing that could improve an individual's decision. The "group" actively suppresses opposing viewpoints, and in some extreme cases, isolates itself from outside influences. So in summary, as a result of people trying to cooperate and make decisions together, a biased environment is created where individuals are less likely to critically question what most of people in their group are proposing. This combination of people wanting to cooperate and not criticising each other makes it more likely for faulty decisions to be made. Groupthink can influence political and corporate decisions, which can result in an enormous waste of human and material resources. It can significantly reduce the efficiency of large groups working on the same project, such as a forecast.

There are ways in which the detrimental effects of groupthink could be swapped for the wisdom of the crowds[36]. This would imply that predictions would be based on the opinion of a group of individuals rather than that of a single individual. The important factor here is that individuals would first make their decisions separately, and would then come together as a group. This would reduce the initial influence a group could have on an individual's decision. Thus, even though groups can cause problems for making predictions, there are certain ways of designing group collaboration that overcome these shortcomings and lead to better predictions.

---

[33] Cass R Sunstein and Timur Kuran, 'Availability Cascades and Risk Regulations' (1999) 51 Stanford Law Review 683.
[34] Marlene E Turner and Anthony R Pratkanis, 'Twenty-Five Years of Groupthink Theory and Research: Lessons from the Evaluation of a Theory' (1998) 73 Organizational Behavior and Human Decision Processes 105.
[35] Marlene E Turner and Anthony R Pratkanis, 'A Social Identity Maintenance Model of Groupthink' (1998) 73 Organizational Behavior and Human Decision Processes 210.
[36] Sheng Kung Michael Yi and others, 'The Wisdom of the Crowd in Combinatorial Problems' (2012) 36 Cognitive Science 452.

III.V. CASE STUDY: BRIDGEWATER ASSOCIATES ALGORITHMIC DECISION MAKING – THE DOT COLLECTOR

Ray Dalio, founder of the investment firm Bridgewater Associates has long been interested in how to make decision-making as algorithmic as possible[37]. In order to do so, he has created programs that aim to eliminate both individual-level and group-level biases. One such example is The Dot Collector, that he presented in his recent TED talk[38]. The Dot Collector collects the views about certain situations from stake-holders and other staff alike, in order to obtain a large diversity of opinions. Employees can also rate each-other's attributes. If someone thinks something about another person's opinion they can convey their assessment by noting an attribute and providing a rating from 1-10. People can also elaborate what their thinking behind the rating was. This tool helps people express their opinions and also separate themselves from their opinions by seeing other people's opinions. This allows people to realise that their opinion is just one of many, and makes them question whether they are right. This can lead to a change in perception, which shifts the group's conversation from arguing over opinions to figuring out objective criteria for determining which opinions are the best. All of these opinions are processed by algorithms so that Bridgewater Associates can know what people are like. This is useful because it allows Bridgewater to know what responsibilities to give them and to weigh their decisions based on people's merits, also known as their believability. When a vote is made about a certain decision, believability is used to weight what decision outcome should be taken into action. Decisions are not made based on democracy, but based on algorithms that take people's believability into consideration. This reduces the impact of individual level false-beliefs and biases. It also makes collective decision-making better than individual decision-making. This approach requires what Ray Dalio likes to call "radical truthfulness and radical transparency"[39]. This is just one example of how group-level decisions can be leveraged for better decision-outcomes.

---

[37] Ray Dalio, *Principles: Life and Work* (Simon & Schuster 2017).
[38] Ray Dalio, 'How to Build a Company Where the Best Ideas Win' (2017).
[39] Dalio (n 37).

IV. PREDICTION MARKETS

IV.I. THE DESIGN OF PREDICTION MARKETS

Another institution that has been successful as a method to aggregate individual forecasts without groupthink is prediction markets. These are markets in which people can buy and sell contracts that have a payoff dependent on whether some future event happens. For instance, people might buy and sell a contract that will pay £1 if Donald Trump is the president of the United States on the 21st of January 2021 (the day after inauguration day), and pay nothing if he is not. The price of this asset will be determined by supply and demand, as in any other financial market, but in a prediction market the price of the asset will reflect the probability of the event happening. If people think that it is very likely that Trump will become president, they will be willing to pay close to £1 to buy the asset, and demand will be high, while if people think it is unlikely but possible they would only be willing to pay a small amount, close to £0, and demand will be low. Under some conditions[40], we can interpret the price of the asset as the aggregate estimate of the probability that the event will happen; if the asset described can be bought for £0.30, then the market participants as a whole believe that there is a 30% chance that Trump will be president in 2021. A prediction market is similar to a British bookmaker, but whereas odds at bookmakers are set so as to maximise profits for the bookmaker, the purpose of the prediction market is to elicit accurate predictions, and so the price of assets is only determined by supply and demand, which in turn are determined by the predictions of the market participants.

Prediction markets may be designed to use real money or play money. Real money markets are just as described above; anyone can buy and sell assets in unlimited amounts using their own money, and the asset will pay out some real money if the corresponding event happens. In a play money market, anyone who wants to participate in the market is given some initial sum of imaginary money, and then is able to use that 'money' to buy prediction market assets. Assets return a fixed amount of this imaginary money. The reward for accurate predictions here is not money, but just a higher point score after a period of time; there may be a leaderboard so that accurate predictors can be rewarded with more prestige. The prices for assets (in terms of play

[40] Charles Manski, 'Interpreting the Predictions of Prediction Markets' (2004).

money) are still set by supply and demand, and we can still interpret these prices as probabilities of the event happening.

Some prediction markets may allow people to make predictions about combinations of events; these are called combinatorial prediction markets. For instance, a basic prediction market might allow you to make a prediction about whether Trump will win Pennsylvania at the next election, while a combinatorial market would also allow you to buy and sell assets that pay off based whether he wins an arbitrary combination of states - e.g. one that pays off if Trump wins Pennsylvania AND Ohio AND Florida. A problem with these markets is that there often won't be many people who want to make a prediction about that particular combination of states - and so there won't be enough people for supply and demand to accurately set the price. As a result, combinatorial markets instead work using a centralised market maker or algorithmic market making method[41]. This market maker offers a price for assets linked to any arbitrary combination of events, choosing this price using a scoring rule that takes into account the price of similar assets. So if someone buys a large number of the assets for Pennsylvania AND Ohio AND Florida - indicating that they think the probability of Trump winning all these three states is actually higher than the previous price offered by the market maker - then this will raise the probability of Trump winning any one of these three states, so the prices in the individual markets for Pennsylvania, Ohio and Florida will also increase. The most commonly used market scoring rule has been described by Robin Hanson[42]. However, combinatorial prediction markets have not been used very heavily in practice, partly because they are more complicated to run and to understand.

The format of questions is another important design consideration. In Section I, we mentioned binary and continuous variable predictions, as well as the intermediate ranked categorical variables. Many prediction markets operate questions across all of these formats, but their choice for each specific question is quite sensitive; in particular, framing biases can be introduced by the market's organizer. Additionally, the exposures of binary and continuous variables to low-probability high-impact events are different, as well as their mathematical

[41] Kathryn Blackmond Laskey and others, 'Graphical Model Market Maker for Combinatorial Prediction Markets' (2018) 63 Journal of Artificial Intelligence Research 421.
[42] Robin Hanson, 'Combinatorial Information Market Design' (2003) 5 Information Systems Frontiers 107.

tractability[43]. This is an intrinsic limitation to the questions proposed by prediction markets, and hence to the accuracy of their forecasts.

So there are various choices to be made in the specific design of prediction markets, but the basic idea is the same: trade in assets whose values depend on future events is used to calculate a forecast of the probability of the events happening, based on all trades in the market.

IV.II. ABSENCE OF GROUPTHINK: EFFICIENT MARKETS HYPOTHESIS AND PREDICTION ACCURACY

There are various aspects of the design of prediction markets that would lead us to expect them to be effective. Firstly, as prediction markets respond to the decisions of large numbers of individuals taking part in the market, they are able to make use of the wisdom of crowds phenomenon to produce more accurate predictions. Groups tend to be wiser than individuals when the relevant information is widely dispersed among them, and no one person or group can have access to all the relevant information. Under these circumstances, every individual who takes part in the prediction market will be bringing some of their own information into the prediction, and so the global prediction will take into account more information than any individual's prediction could. We can think of the price mechanism used by prediction markets as just a particular algorithm that can be used to aggregate the predictions of all the participants in the market, taking into account their confidence in their predictions (this interpretation is discussed by Snowberg *et al.* (2013)[44].) This argument only holds when information is dispersed, which sheds light on the situations in which prediction markets will be most useful. If it is possible to consolidate all relevant information about the event into a single source – which is often the case for natural events, such as earthquakes – then the prediction of a crowd, or a prediction market, will not be much more accurate than an individual prediction made by the person or organisation with that information. However, if the relevant information is widely dispersed and could not conceivably be consolidated into a single source – which is often the case for social or political events, such as elections or wars – then prediction markets and other methods that leverage the wisdom of crowds will have a substantial advantage.

---

[43] Herman Chernoff, 'A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the Sum of Observations' (1952) 23 The Annals of Mathematical Statistics 493.
[44] Erik Snowberg, Justin Wolfers and Eric Zitzewitz, 'Prediction Markets for Economic Forecasting', *Handbook of Economic Forecasting*, vol 2 (2013).

Markets and prices have often been thought of as a particularly effective way to aggregate dispersed information. This was the basis of Hayek's defence of free markets and criticism of centrally planned economies[45], which pioneered this view of markets as carrying information. Subsequently, Fama and others developed the Efficient Market Hypothesis, which posits that the price of financial assets fully reflects all available information about the value of the asset[46]. In a practical sense, the argument is that anyone who has information about an asset that suggests its price is currently too low relative to its value has a financial incentive to go and buy it as soon as possible – before anyone else is able to do so – thereby increasing the price and incorporating their information into the asset's price. In the context of a prediction market, the value of the assets is determined entirely by the probability that the related event will happen. Consequently, if prediction markets are efficient, we would expect them to be able to incorporate all the dispersed information about the probability of events, and to react to new information affecting that probability more rapidly than other prediction methods. As an example, Figure 3 below shows how the InTrade prediction market was able to determine within minutes that Osama Bin Laden had been killed based on (reliable) social media information, before this news was officially announced by President Obama or reported by mainstream media outlets; the price for this asset reached \$100, indicating a 100% probability, well before Obama's address. Another example is from the Challenger space shuttle disaster[47]. There were four major private contractors involved with the shuttle; the source of the fault was unknown at the time of the disaster, but the Rogers Commission, which was set up to investigate it, concluded after several months that the fault was caused by the company Morton Thiokol. However, on the day of the disaster itself, the stock price of Morton Thiokol declined faster than that of the other contractors, and while their stock prices rebounded after some time, Morton Thiokol's did not. The market was apparently able to identify Morton Thiokol as the culprit for the disaster months before any information confirming this became public.

[45] FA Hayek, 'The Use of Knowledge in Society' (1945) 35 The American Economic Review 519.
[46] Eugene F Fama, 'Efficient Capital Markets: A Review of Theory and Empirical Work' (1970) 25 The Journal of Finance 383.
[47] Michael T Maloney and J Harold Mulherin, 'The Complexity of Price Discovery in an Efficient Market: The Stock Market Reaction to the Challenger Crash' (2003) 9 Journal of Corporate Finance 453.

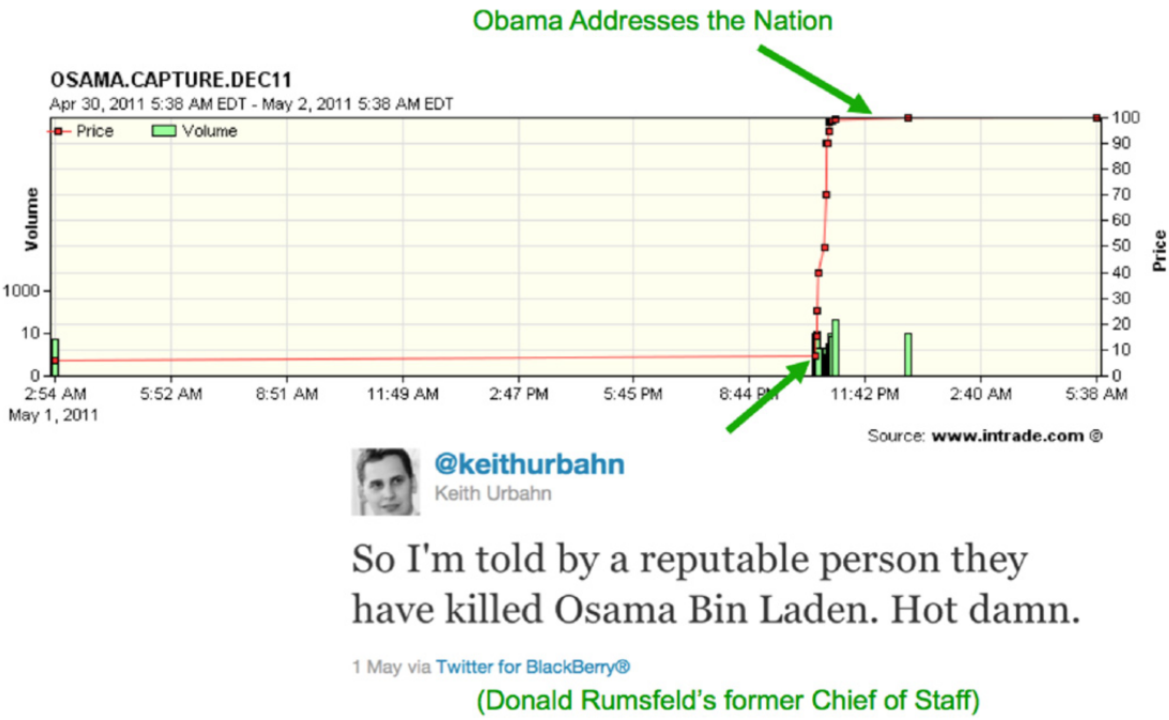

Figure 3: Response of prediction markets to capture of Bin Laden; reproduced from Snowberg *et al.* (2013)[48]

These arguments are typically based on markets where real money is at stake. However, as discussed above, many prediction markets are based on play money. Nevertheless, these markets may still retain some of the benefits of real money prediction markets. Play money markets often publish rankings of participants, and the bragging rights and reputation produced by this ranking function as an alternative incentive to perform well. Some play money markets may provide cash or other prizes to some individuals near the top of the ranking after some period of time, which provides further incentives to predict accurately. Play money markets are typically less costly to run and less strictly regulated; furthermore, they alleviate a problem in real money markets, which is that richer people can afford to trade larger volumes in the market, and so have a disproportionate effect on the market price relative to their actual confidence. However, markets where the reward is real money rather than play money provide stronger incentives for people to put effort into making accurate predictions. Empirically, these factors may balance each other out; a comparison of forecasts for sporting events from the real

[48] Snowberg, Wolfers and Zitzewitz (n 44).

money market TradeSports and the play money market NewsFutures did not find a significant difference in accuracy between the two different types of market.[49]

Incentives, whether financial (in real-money markets) or reputational (in play-money markets), play an important role in making prediction markets an effective mechanism. Any participant in a prediction market has a direct incentive to make unambiguous and verifiable predictions about the world, to publicly and truthfully share those predictions (by using them to decide when to buy and sell prediction market assets), to update those predictions when new information comes to light rather than holding on to beliefs that are inconsistent with new evidence, and to seek out information that will improve the accuracy of their predictions; other mechanisms for forecasting do not provide such relevant incentives. In short, prediction markets encourage participants to put their money where their mouth is, and the quality of the predictions that we can get by aggregating individual predictions is improved as a result. In addition, prediction markets especially reward people who are able to make forecasts based on some under-appreciated sources of information. People who just make forecasts based on easily available public information are unlikely to make a lot of profit out of prediction markets, because the efficient market hypothesis states that this information will be incorporated into the market price very quickly. However, someone with a private source of information will be able to make substantially more profit because most other investors do not have access to the same information, and cannot make the same trades to exploit it. In this way, prediction markets provide a particular incentive to seek out diverse and unusual sources of information, which is likely to increase the total information that the aggregate prediction takes into account relative to other forecasting mechanisms.

IV.III. PRACTICAL APPLICATIONS: PAST AND FUTURE

There are several ways in which the theoretical benefits of prediction markets can be realised in practice. The various benefits discussed in the previous section have been a driving force for the implementation of prediction markets.

One of the most common, and also one of the earliest, practical applications of prediction markets is in predicting election results. Betting on political outcomes has become far more

[49] Emile Servan-Schreiber and others, 'Prediction Markets: Does Money Matter?' (2004) 14 Electronic Markets 243.

sophisticated since the papal selection in 16th century Italy[50] with the advent of political stock markets. The contracts in these prediction markets are based on the outcome of a given election. The relative prices of the various contracts give an indication of the likely probabilities for victory for the candidates. However, various contracts can exist other than those simply predicting the success of a given candidate. Intrade, a former web-based market, has offered contracts on voter turnout, Vice-Presidential candidates and control of the House and Senate among others[51].

The Iowa Electronic Markets (IEM) has been one of the most widely used markets in recent times. Run by the University of Iowa Henry B. Tippie College of Business, primarily as a research tool, the IEM has been predicting election outcomes since 1998. As one of the most popular markets, their performance can serve as a benchmark to judge the quality of political stock markets in general as a tool for predicting election outcomes. Jacobs (2009) compared the IEM against opinion polls in the 2008 US election. By comparing the vote-share market as opposed to winner-take-all contracts (as "data from winner-take-all contracts for individual candidates or parties cannot be judged for accuracy against the outcome of the election"), Jacobs found that the IEM was closer to the actual election outcome 66% of the time[52].

Prediction markets can also be used to aid in making policy decisions. When used in this capacity, they are commonly referred to as Decisions Markets. Hanson (1999) described markets of this nature that utilise market information to make decisions as "decision support systems"[53].

Berg & Rietz (2003)[54] explored the potential use of the IEM as a tool to choose the Republican candidate in the 1996 US election. Conditional contracts based on future vote share (e.g. contracts paid out "$1 times the Democratic nominee's vote share conditional on Robert Dole being the Republican nominee") were issued which offered information as to how each candidate would fare against Bill Clinton. The value of these various contracts gave information about how each candidate would fare against Clinton. The Republican Party went with the

[50] Paul Rhode and Koleman Strumpf, 'Historical Political Futures Markets: An International Perspective' (2008).
[51] Vincent Jacobs, 'Prediction Markets: How They Work and How Well They Work' (Catholic University of Louvain 2009).
[52] ibid.
[53] Robin Hanson, 'Decision Markets' (1999) 14 IEEE Intelligent Systems and their Applications 16.
[54] Joyce E Berg and Thomas A Rietz, 'Prediction Markets as Decision Support Systems' (2003) 5 Information Systems Frontiers 79.

Primary Results and chose Robert Dole who went on to lose the election. The markets predicted that Colin Powell would have been the far stronger candidate. As the likelihood of him being the Republican Candidate fell, the probability of Bill Clinton winning the election grew. If the markets had been used as a tool to facilitate the process of selecting the candidate, there might have been a very different election outcome.

A very theoretical and idealised application of Decision Markets is their potential use in a futarchy. Proposed by Hanson, he believed that the forecasts made by the market could be used to support all kinds of policy decisions. In our current form of governance, policy makers frequently ignore the advice of relevant experts, and the advice itself is of a varying degree of quality. The combination of these factors leads to poor decision-making. A futarchy is a form of government where "democracy would continue to say what we want, but betting markets would say how we get it"[55]. Although the chances of such an idea being realised are currently very slim, Hanson does drive the conversation towards looking at further applications of Decision Markets. One such application, the Policy Analysis Market, will be explored in a later section.

The markets discussed so far, such as the IEM, have relied on publicly available information with a wide range of traders. However, corporations utilise a very distinct variant which aggregates private information instead. Internal Prediction Markets are used to make business forecasts. The nature of these markets allows employees to share bad information which they would otherwise not feel comfortable doing. An example would be if an employee believes that a particular project will overrun or be over budget yet fears that voicing these concerns to senior management would harm their career progression.

One such application is the internal market run by Hewlett Packard (HP) from 1996 to 1999 to forecast printer sales. Chen & Plott (2002)[56] found that not only did the market outperform the official forecasts, it was even able to accurately predict whether the official forecasts would be too high or too low. Another example is the internal markets run by Google. Cowgill et al

[55] Robin Hanson, 'Shall We Vote on Values, But Bet on Beliefs?' (2013) 21 Journal of Political Philosophy 151.
[56] Charles Plott and Kay-Yut Chen, 'Information Aggregation Mechanisms: Concept, Design and Implementation for a Sales Forecasting Problem'.

(2009)[57] found that although an optimistic bias was initially present, with time the market calibrated and allowed accurate forecasting of questions such as whether a certain project would be completed on time.

Corporations can also rely on public markets in order to facilitate decision-making. Berg et al (2007) explored how prediction markets could have been practically used to forecast the post-IPO share price of Google and inform the valuation of the company before it was listed. They concluded that the public prediction markets could be successful at this and were accurate even before some information related to the IPO, such as the number of shares, became public. Hypermind is another example of a public market being utilised to aid corporations. Hypermind runs a play money market with cash prizes paid for by sponsors. The traders are members of the public chosen by Hypermind themselves and the questions at any given time (e.g. When will Theresa May trigger Article 50?) are chosen by the sponsors.

IV.IV. ACHIEVEMENTS AND FAILURES

Prediction markets have had multiple successes in the real world, as well as in laboratory experiments. As discussed earlier, the Iowa Electronic Markets have been a highly accurate tool for predicting the results of elections. Another example of a successful prediction market is Hypermind which has also proved to be an invaluable tool for companies to utilise their forecasts to aid their planning and long-term strategies. The successful application of prediction markets also extends to internal markets used to aggregate private information. The markets run by HP and Google in the previous section offer a snapshot of the potential benefits afforded by Internal Prediction Markets.

However, there are also several downsides to prediction markets. One of the biggest risks is their susceptibility to manipulation. Manipulators are traders present in the market who seek to alter prices for either their own profit, or to change the forecast to align with their interests. The two methods of gauging the resilience of prediction markets to manipulators are to either study the effects in a controlled experiment or to observe the effects of manipulators in a real-life market. An example of the first are the series of experiments ran by Deck and Porter[58] to study

[57] Bo Cowgill, Justin Wolfers and Eric Zitzewitz, 'Using Prediction Markets to Track Information Flows: Evidence from Google' (2009).
[58] Cary Deck, Shengle Lin and David Porter, 'Affecting Policy by Manipulating Prediction Markets: Experimental Evidence' (2013) 85 Journal of Economic Behavior & Organization 48.

manipulators. They found that although manipulators can affect the price, which can in turn lead to poor forecasts, manipulators do not have an effect on bid and ask queues. This extra information included in the market but not in the prices can still be used by forecasters. Rhode and Strumpf[59] examined the resilience of historical political betting markets to manipulators. They concluded that "in almost every speculative attack, prices experienced measurable initial changes. However, these movements were quickly reversed and prices returned close their previous levels". In spite of these studies, concerns remain about the risk posed by manipulators and research is ongoing into this.

Additionally, well-intentioned forecasters can also introduce new biases into their predictions, especially depending on the format and phrasing of specific questions. For example, it is easy to mistakenly equate the payoff of the question too strictly for the probability of the forecaster event, which is potentially dangerous in fat-failed domains. Attempts to hedge continuous-variable bets with binaries creates an important risk for attribute substitution [60].

Although the internal markets examined earlier were illustrations of successful instances, there has been an extremely limited uptake of prediction markets into corporations. There are two fundamental reasons for this. The first, is that the advantages of forecasting with the use of prediction markets break down in certain circumstances. In situations where there is a lack of historical data (e.g. a product being deployed in a new market), prediction markets excel. However, where there is previous data available, the advantage of prediction markets relative to statistical modelling often disappears.

The second reason is related to how internal prediction markets conflict with current corporate culture. Forecasting within corporations can be derailed by political influences, and used as a tool for reasons other than gaining accurate information about the future, such as purposefully advancing pre-conceived ideas or policies. A prediction market by its very design is fully transparent: every single trader participating in the market from the factory floor to the boardroom will be able to observe the same information. Corporate leaders often do not desire this level of transparency because it will conflict with their own personal goals which are not

[59] Paul W Rhode and Koleman S Strumpf, 'Historical Presidential Betting Markets' (2004) 18 The Journal of Economic Perspectives 127.
[60] Nassim Nicholas Taleb and Philip E Tetlock, 'On the Difference between Binary Prediction and True Exposure with Implications for Forecasting Tournaments and Decision Making Research' [2013] SSRN Electronic Journal.

always aligned with that of the company as a whole. As Emile Servan-Schreiber, co-founder of Hypermind, put it, the culture problem is a "generational struggle"[61]. According to him, the conflict has arisen due to the current generation being used to and comfortable with sharing their opinion and having information freely available. The existing power structure, in general, are not only unaccustomed to knowledge sharing, but they also have no interest in it.

IV.V. CASE STUDY: POLICY ANALYSIS MARKET

Policy makers have long sought ways to improve the quality of decision-making. As the interest in prediction markets began to pick up towards the end of the 20th century, the US government began taking an active interest in their potential applications. The Policy Analysis Market (PAM) was a project of the United States Defence Advanced Research Project Agency (DARPA). A revolutionary concept, it eventually met an untimely end due to huge public backlash. Denounced by politicians and the public alike, it was cancelled a day after it was launched[62] and led to the resignation of Admiral John Poindexter.

The value of prediction markets became apparent to DARPA during the Clinton administration. DARPA began funding research on prediction markets and gave a call for proposals in 2001 titled "Electronic Market-Based Decision Support". Net Exchange's proposal would go on to secure the most funding, and would later be named the Policy Analysis Market.

According to Robin Hanson, the system architect of the project run by Net Exchange, the team reasoned "that the cost to create markets does not depend much on the topic, but the value of estimates varies enormously with the topic; thus, the greatest benefit relative to cost would come from the highest value estimates."[63] Forecasts aiding defence policy decisions were seen as the most valuable and so this would be the main focus of PAM. The designers planned to be able to use PAM to forecast the relationship between US policies and political instability around the world, amongst other issues.

---

[61] In conversation with authors

[62] Julian Borger, 'All Bets Are off on Terror, Rules Pentagon' *The Guardian* (30 July 2003) <https://www.theguardian.com/world/2003/jul/30/usa.julianborger>.

[63] Robin Hanson, 'The Policy Analysis Market (A Thwarted Experiment in the Use of Prediction Markets for Public Policy)' (2007) 2 Innovations: Technology, Governance, Globalization 73.

The designers later refined PAM and focused instead on eight Middle Eastern nations. Traders in the market would be able to create prices to estimate five parameters for each of the eight nations. These were: military activity, political instability, economic growth, US military activity and US financial involvement. In addition to these, other parameters were added bringing the total number of base parameters to be estimated up to a few hundred. Although it was intended that PAM would be run between government agencies, strong legal barriers meant that PAM was opened up to the public in order to have enough traders. The potential uses of PAM were numerous. For example, Hanson suggested that it could have been used to predict the "local and global consequences of invading Iraq".

However, the premature demise of PAM means that there is no available data for analysing the effectiveness of the project. As stated above, PAM was shut down due to a strong public backlash, which began when Democrat Senators Ron Wyden and Byron Dorgan held a press conference on July 28th and lambasted the plan.[64] This was followed the next day by Senate Democratic Leader Tom Daschle who took the issue to the floor of the Senate and branded PAM "a plan to trade in death".[65] Furthermore, the association of PAM with DARPA executive John Poindexter, the man who was convicted in the Iran-Contra scandal, also tainted the public image of the project.

The failure of the project highlights the importance of managing public opinion. In the case of PAM, although the claims made against it were flimsy and easily countered, the weight of public perception was too great. If more effort had been made to manage the political fallout and brand the project in a more acceptable way for the public, PAM might have lasted and been a source of valuable information as to the effectiveness of prediction markets in practice, not to mention the benefit to the US government.

IV.VI. COMPARATIVE EFFECTIVENESS

Formal comparative studies of prediction markets and their effectiveness are useful for understanding when we can expect prediction markets to improve on other forecasting methods. An alternative prediction method for elections has been the forecasts of FiveThirtyEight and similar websites, which aggregate opinion polls using a (somewhat opaque)

[64] Ron Wyden and Byron Dorgan, 'Wyden, Dorgan Call For Immediate Halt to Tax-Funded Terror Market' (*U.S. Senate Press Release*, 2003) <http://wyden.senate.gov/media/2003/print/print_07282003_terrormarket.html>.
[65] BBC News Americas, 'Pentagon Axes Online Terror Bets' *BBC News* (29 July 2003).

model to provide running predictions of the results of elections. Rothschild (2009)[66] compared FiveThirtyEight's forecasts in the 2008 US election for the vote of each state in the Presidential election, as well as in the Senate, to forecasts from the public, real-money prediction market Intrade. Comparing the raw forecasts of the two models, the prediction market performed as well as FiveThirtyEight's forecasts in the first half of the election campaign, but fell behind in the second half. However, when a standard transformation was applied to correct for the known favourite-longshot bias of prediction markets, Intrade's performance improved; the debiased Intrade performed significantly better than FiveThirtyEight in the first half of the campaign, and matched its performance in the second half. Furthermore, a test used by Rothschild showed that there was little evidence that using the FiveThirtyEight forecast provided any extra information than the debiased InTrade. So in this case, using prediction markets naively is not an improvement on other forecasting methods, but if we account for the known biases of prediction markets, they do appear to contain more information and be somewhat more accurate in predicting election results.

FiveThirtyEight's forecasts are based on centrally collected information, and do not make use of the wisdom of the crowd. The fact that they performed at the same level as prediction markets towards the second half of the campaign suggests that in the case of elections, most of the relevant information can be collected centrally, largely in the form of opinion polls. Although, prediction markets' advantage in the earlier stages suggests that polls do not provide the same information towards the start of a campaign. A paper by Atanasov *et al.* (2017)[67] instead compares prediction markets to the averages of the predictions of many forecasters – this is a crowd-based method. This research was part of the IARPA/Good Judgement study (discussed in Chapter 1), and so focused on geopolitical events. People who joined the project were randomly assigned to make forecasts either individually (with no communication with other forecasters), as part of a team of about 15 members (where each individual could communicate with the other members of the team, but not with people from other teams), or as part of a play-money prediction market (where there was no direct communication between individuals, but they could see changes in the aggregate forecast by observing changes in the

[66] David Rothschild, 'Forecasting Elections' (2009) 73 Public Opinion Quarterly 895.
[67] Pavel Atanasov and others, 'Distilling the Wisdom of Crowds: Prediction Markets vs. Prediction Polls' (2017) 63 Management Science 691.

price). All the forecasts were probabilistic forecasts of binary outcomes. The forecasts of the individuals and those of the different teams were then aggregated to produce two additional aggregate forecasts, one for the individuals and one for the teams, which could be compared to the prediction market forecasts. These non-market aggregate forecasts are known as prediction polls.

The authors found that, when prediction markets were compared to prediction polls – either of individuals or of teams – that combined the different forecasts by simply calculating their mean, prediction markets performed significantly better, having a significantly lower Brier score. They attributed this in part to three consistent problems with using a simple average in prediction polls: failing to weight recent predictions (based on more up-to-date information) more heavily, failing to weight individuals based on their accuracy, and consistent underconfidence, where all predictions were too close to 50%.

However, all of these problems can be corrected for while still using prediction polls, simply by using a more complicated algorithm to combine the different predictions. The algorithm used weighted forecasters based on their Brier score across previous questions, excluded all but the 20% most recent forecasts, and recalibrated the aggregate forecasts to push the values towards the extremes. Under this aggregation algorithm, the aggregate forecasts of both types of prediction poll improved; the team forecasts became significantly more accurate than prediction markets, while the individual forecasts were as accurate as prediction markets. Applying similar transformations to the prediction market forecasts did not improve their accuracy, and in some cases worsened it, so the best obtainable forecast from team prediction polls was better than the best obtainable forecast from prediction markets.

Further analysis of the data showed that the advantage of team prediction polls over prediction markets was solely in relatively long-term questions, that were open for a longer time period than the median (105 days); for questions with a shorter time horizon, team prediction polls and prediction markets were equally accurate. This time horizon difference has been attributed to the opportunity cost of long-term predictions in markets[68]. In short, when people discount the future and value money today over money in the future (even virtual money), they will tend

[68] Lionel Page and Robert T Clemen, 'Do Prediction Markets Produce Well-Calibrated Probability Forecasts?' (2013) 123 The Economic Journal 491.

to prefer making a series of short-term predictions in prediction markets to having their money locked up in a long-term prediction. Figure 4 illustrates the difference in calibration based on time horizon; a perfectly calibrated forecast should be on the 45 degree line in this figure. Prediction polls do not face this problem, as there is no opportunity cost of making a forecast for the long term relative to the short term.

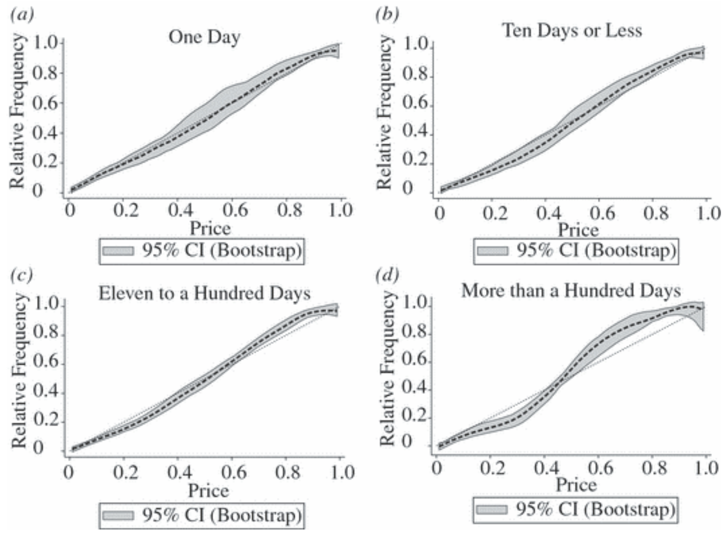

Figure 4: calibration of prediction market forecasts by time horizon of the forecast, reproduced from Page and Clemen (2013)

These comparative studies provide positive results for crowd-based forecasting methods, at least in the domain of politics and geopolitics; InTrade's forecasts outperformed FiveThirtyEight's, and crowd-based methods were the most successful in the GJP. However, the Atanasov *et al.* paper suggests that markets are not the only successful way to elicit and aggregate the forecasts of a crowd, and that another aggregation algorithm may work at least as well. The communication between forecasters facilitated by team prediction polls appears to improve the accuracy of forecasts, as information can be shared and discussed between forecasters, meaning that each forecaster can directly see perspectives that they might not have otherwise considered. The potential for groupthink seems not to have offset these benefits of teams, although this may be because the participants in the GJP were particularly good at avoiding groupthink. It might be possible to combine this with prediction markets by encouraging teams of people to trade collectively on prediction markets, rather than have individuals trade.

Prediction polls have particular advantages over prediction markets in situations where the number of forecasters is small, or the time horizon of the forecast is large; conversely, prediction markets may perform as well or better when there are many forecasters and the time horizon is relatively short.

Another factor affecting the choice between prediction markets and prediction polls is that prediction markets aggregate forecasts automatically, instantly and transparently, while prediction polls require a particular aggregation algorithm to be chosen, and the aggregate forecast is not necessarily publicly available. Prediction polls therefore give the managers of the prediction system more control over how the forecasts are aggregated. If it is unproblematic for all participants to know the market prediction at all times, then prediction markets may be more suitable, but if there are advantages to keeping the aggregate prediction private, prediction polls may be more suitable. Prediction polls may prove more suitable for private-sector settings than prediction markets, since there are often reasons why managers may not want all employees to know the current aggregate prediction.

V. LONG-TERM FORECASTING

V.I. INTRODUCTION

Given the range of methods discussed for obtaining reliable forecasts, we may now ask how far into the future these techniques allow us to predict. Long term forecasting is essential for all time-sensitive matters, where decisions, preparations and actions have to be made as far in advance of an event or deadline as possible. For instance, accurate forecasting allows for an optimized allocation of government or institutional resources. Furthermore, some issues are intrinsically long-term, hence long-term forecasting is the only common-sense solution to tackling this problem with full efficiency. It is important to note that "long-term" refers to widely different timescales depending on context. This term corresponds to decades for population projections, months for political forecasts or several days for weather forecasts.

Forecasters have a long track record of successful long-term predictions when dealing with causal models[69], such as predicting climate change or planetary motion[70]. In such models, all of the causal mechanisms of a system are well defined, and hence this allows for computational modelling. The issue in long-term forecasting comes from making long-term predictions in complex systems where small and/or random effects can propagate. In such systems, models must take into account uncertainties in an event or phenomenon. These models take into account random variables into the model of an event or phenomenon. Unless the models are extremely precise, the uncertainty that is introduced by the random variables will compound over time, leading to unreliable or unhelpful forecasts concerning possible future scenarios.

There have been many failures in long-term forecasting. Examples include predictions on GDP growth[71], political elections[72] and energy consumption over long time frames. In most cases, even with the best methods and when a forecaster has an economic interest in making the most accurate predictions, their correctness is sub-par. On the other hand, there are still many gaps in the literature of this field. This is mostly because long-term forecasting lies at the current frontier of research: to date, only few legitimate attempts at making long-term forecasts

[69] Judea Pearl, *Causality*, vol 19 (Cambridge University Press 2009).
[70] Helen Beebee, Christopher Hitchcock and Peter Menzies (eds), *The Oxford Handbook of Causation*, vol 1 (Oxford University Press 2010).
[71] Andrew Ang, Monika Piazzesi and Min Wei, 'What Does the Yield Curve Tell Us about GDP Growth?' (2006) 131 Journal of Econometrics 359.
[72] JE Campbell and JC Garand, *Before the Vote: Forecasting American National Elections* (Sage Publications 2000).

have been produce, and there is a lack of studies on the accuracy of these forecasts. In general, such long-term forecasts either relied on a causal model and aren't relevant to this study, or gave inaccurate predictions. Hence, future research is needed to conclude what factors will improve long-term forecasts.

V.II. DIFFICULTIES: TIME, BIASES, SUDDEN EVENTS

Some of the difficulties of long-term forecasting arise from cognitive biases, while others arise from the difficulties of evaluating of long-term forecasts. When making long-term forecasts, the path to the ultimate future event that we are interested in can often be broken down into a series of smaller milestones. For instance, there is substantial interest in the question of whether Artificial General Intelligence (i.e. one that can replicate all the intellectual capabilities of a human being) will be developed by 2050 for instance. It is possible to consider several milestones that one would expect to reach in particular years if AI were developing at a pace such that AGI would be developed by 2050. Making predictions based on these milestones has the advantage of being verifiable at an earlier date than the overall prediction, and gives more detail about *how* AGI will be developed rather than just *whether* it will be developed. However, the disadvantage is that the probability of AGI being developed along a particular path is lower than the probability that it will be developed at all; from basic probability theory, P(A and B) ≤ P(A). Conjunction bias, a common cognitive bias, means that people often have a problem with understanding this problem. If a scenario described for the development of AGI coincides with how they expect AI to develop over time, they will tend to overestimate its probability relative to more general scenarios that specify fewer details.[73]

For the purposes of formal prediction, however, we need to realise and account for the fact that there is a tradeoff between the detail of a prediction and the probability of that prediction being correct. Predictions that are more vague (but still verifiable) have a higher probability of being accurate than more fully specified predictions. In some cases, particular intermediate milestones may be necessary or sufficient for the ultimate event, and in these cases it may be useful to focus our predictions on these intermediate milestones. In general, however, there are a vast number of possible paths into the future, which create a 'combinatorial explosion'.

[73] Amos Tversky and Daniel Kahneman, 'Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment.' (1983) 90 Psychological Review 293.

Several cognitive biases that we have discussed previously are amplified when the probabilities of events are small. People are generally bad at evaluating very small probabilities (below 1%), and particularly differences between them; instinctively we interpret the odds "one in a million" and "one in a billion" as both just meaning "very unlikely," even though the former is far more likely than the latter. This means that, when estimating the probability of a very unlikely event, people tend not to assign the correct probability. Unlikely events that are relatively easy to imagine, or ones which have a particularly detailed and vivid description associated, or events where the alternative is less specifically described, will be overweighted so that people consistently overestimate the probability of these events; this is linked to the availability heuristic, which dominates the effects of more rational estimation of the probability at low levels. On the other hand, events that are more difficult to imagine – perhaps because they have never happened yet – will generally be underweighted relative to the true probability. This is described as neglect.

These biases help to explain the planning fallacy, which is the particular tendency for people involved in a decision to be overly optimistic about the probability of success on time; because success is usually more clearly defined than failure, especially for the decision-makers, the probability of success will be overestimated. This is a problem if, for instance, we ask people working in AI about the likely timetable for its development. The overweighting of more detailed and vivid events is also linked to conjunction bias; people do not take account of the fact that a clearer specification of the path to a long-term future event reduces the probability that this path will be realised, and thus ignore the tradeoff between probability and detail. Finally, when small probabilities are involved, we need to take account of the probability that our approach to forecasting the question – the assumptions or information on which we based probability estimates – was fundamentally wrong. The chances of this may be larger than the small probability of the event itself, and if the prediction is very sensitive to the forecasting approach, then this effect will probably dominate the actual estimate of the event's probability.

However, methods to reduce the effect of similar biases were discussed earlier in the paper, for instance through training or through a formal aggregation mechanism such as a prediction market. In long term forecasting, problems are further exacerbated by a lack of feedback from events to the predictions. Good short-term forecasters, or forecasting methods, are able to learn from mistakes they make by looking for patterns in the occasions where their predictions are wrong, and correcting for any systematic biases in their prediction. With long-term forecasts, this feedback loop is weaker, or even non-existent. The event may take years, or decades, to be realised, and forecasters thus have to wait a long time before learning how accurate their prediction was. This makes it very difficult to evaluate particular long-term forecasts or forecasting methods, relative to short-term methods. We have previously emphasised the importance of assessing and scoring predictions to compare forecasting techniques; unfortunately, for long-term forecasting, this is frequently not possible, which makes it difficult to improve methods over time. This applies both to the accuracy of particular forecasts and the overall calibration of a forecasting method. Prediction markets are likely to fail for long-term events because, as discussed previously, people discount money that they will receive in the future, and would only be willing to take part in the market if they were guaranteed a high amount of interest on the money they invested, in the event that they won. The Delphi method has also been used for long-term forecasting, but the individual experts are still likely to succumb to the cognitive biases that make long-term forecasting difficult. Methods combining several forecasting techniques, such as multi-criteria assisted Delphi studies [74], have been proposed to make use of complementary strengths, in order to overcome the unique difficulties of long-term forecasting, on top of the overall challenge of forecasting.

V.III. OVERCOMING THE DIFFICULTIES OF TIME: SCENARIO GENERATION AND DEBIASING

A particularly tempting solution to the loss of accuracy for long-term forecasts is to aggregate increasingly large numbers of predictions. The hope is that the wisdom of experts, or of the crowd (or perhaps even both), accrues forecasting accuracy to overcome the difficulties mentioned previously. And while it may be counterintuitive that infinitely combining very inaccurate forecasts will consistently lead to more accurate aggregated predictions, the wisdom of the crowd phenomenon has been repeatedly demonstrated even in such extreme circumstances[75].

---

[74] Ahti Salo, Tommi Gustafsson and Ramakrishnan Ramanathan, 'Multicriteria Methods for Technology Foresight' (2003) 22 Journal of Forecasting 235.
[75] James Surowiecki, *The Wisdom of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations.* (Doubleday & Co 2004).

Moreover, aggregating a large number of individual judgements could help forecasters deal with the combinatorial explosion problem in another way, namely to focus on broad characteristics rather than very specific predictions to tackle the problem head-on. If multiple reliable forecasts can be made about various broad characteristics, that each hold over a subset of different possible futures, a significant fraction of the tree of possible scenarios may be discarded. Of course, the forecasting difficulty is then moved to the definition and choice of these characteristics. Such a method would also require that the chosen characteristics apply to sufficiently different sections of the tree of scenarios. However, if this can achieve, the combination of these forecasts can then be straightforwardly combined using logical arguments.

In many cases, long-term forecasts will require forecasters to make a series of assumptions, and to consider their predictions in light of these assumptions. Beyond the benefit of eliciting conjectures that could otherwise have remained implicit and led to biases, these can add to the refinement of the models used. Other methods in scenario generation have been developed to aid the systematic discovery of scenarios, with a specific emphasis on avoiding typical cognitive biases[76].

While these may seem like obstacles to the eventual convergence towards an accurate forecast, such hypotheses provide direct advantages to the forecasting framework. Short-term milestones derived from these assumptions are of particular interest, as they can subsequently be used as feedback to the forecasters. This is true even if the milestones have no intrinsic value.

Such methods are especially useful when dealing with technological progress, as they generally provide information at each stage of improvement. Important examples include the Whole Brain Emulation project, which seeks to model the function of the brain one-to-one, i.e. recreate all of the intellectual capabilities of the human brain[77]. As it is currently only a theoretical technology, various methods have been proposed to forecast its development. An important takeaway from this research is the requirement of falsifiable design, which in essence states that designers of theoretical technologies should propose tests to gradually reduce all

[76] Muhammad Amer, Tugrul U Daim and Antonie Jetter, 'A Review of Scenario Planning' (2013) 46 Futures 23.
[77] A Sandberg and N Bostrom, 'Whole Brain Emulation: A Roadmap' (2008).

allowed uncertainties[78]. Other works, such as Age of Em[79], focus on the possibility that an artificial super intelligence is not feasible, but that human minds can be uploaded to computers.

More generally, much attention should be given to debiasing and training experts to avoid the pitfalls discussed above. The Good Judgment Project revealed that the more accurate a forecaster was in general, the better they performed for questions with long timescales. In particular, so-called super-forecasters representing the 10% most accurate participant were found to be more scope sensitive, i.e. better at taking into consideration the time frame of predictions; this observation led to revisions in the GJP's later debiasing experiments[80]. This strongly suggests that the ability to avoid cognitive biases is linked to accuracy of long-term forecasts.

Training is particularly important for long-term forecasting, where the limited amount of relevant information means that eliciting this information fully and in an unbiased way is critical[81]. Most likely, forecasts will depend on expert judgement, such that development of methods that are more effective than expert elicitation may be a suitable complement. Furthermore, obtaining expert judgement is generally cheap. Scalable and inexpensive training is therefore important, to complement these advantages over other forecasting techniques.

Different metrics may also correspond better to the challenges presented by long-term forecasting. For instance, instead of using the Brier score to assess the accuracy of predictions, as discussed previously, the logarithmic scoring rule (as mentioned in Section II.2.) is better suited to long-term forecasts: it places a heavy penalty on failure to predict extremely likely or unlikely events. In that sense, forecasters assessed with a logarithmic rule will be incentivized to minimize the expected surprise[82]. Choosing the relevant scoring metric also has significant broader implications, such as allowing proper calibration of the forecasters to be performed under the conditions of long-term forecasts.

---

[78] N Szabo, 'Falsifiable Design: A Methodology for Evaluating Theoretical Technologies' (2007) <http://unenumerated.blogspot.co.uk/2007/02/falsifiable-design-methodology-for.html>.
[79] R Hanson, *The Age of Em: Work, Love, and Life When Robots Rule the Earth* (Oxford University Press 2016).
[80] Tetlock and Gardner (n 2).
[81] Welton Chang and others, 'Developing Expert Political Judgment: The Impact of Training and Practice on Judgmental Accuracy in Geopolitical Forecasting Tournaments' (2016) 11 Judgment and Decision Making 509.
[82] Gneiting and Raftery (n 8).

V.IV. BLACK SWANS

In his 2007 book, Nassim Taleb forewarned that there are certain highly improbable events that are too important to ignore[83]. He calls such cases Black Swan events. Black swan events are unexpected but extremely high impact events, which can be catastrophic for human civilisation. Due to the devastation that they can cause, Taleb argues that they should not be ignored. Examples include violent acts of terrorism, epidemics and financial crashes.

The issue here isn't that people do not know about these events. Rather, the issue is that they are usually neglected, hence we are not fully prepared for them. They are events which have predicted probabilities below an action threshold[84]. They are usually surprises because the people's predictions were wrong, or because low probability events still occur. More specifically there are "unknown knowns" - events that are a surprise to people that were conducting the risk analysis - and "unknown unknowns" - events that are completely unexpected.

We cannot fully solve the Black Swan problem, but there are partial solutions. One of these is having an explicit measure of impact that makes it harder to neglect events based on probability alone. In general good forecasting practice will make it easier to avoid unknown knowns and at least identify some unknown unknowns. Tools such as scenario planning, used along with components of devil's advocacy and strict dialectical thinking, have been suggested to overcome the weaknesses of many pure forecasting techniques[85].

Although not strictly part of forecasting but rather decision-making, it is important to note another partial solution: making systems "anti-fragile", i.e. creating economic, societal and technological systems that gain robustness in response to shocks and catastrophic events[86]. Under that approach, the question is no longer about simply forecasting such black swans in order to react early to their effects; instead, an entire mechasim is designed or re-designed to respond positively to stress or chaos, nullifying the threat of black swans of a much more generic nature. This is particularly important in "ruin scenarios", in which damage is

[83] NN Taleb, *The Black Swan: The Impact of the Highly Improbable* (Random House Publishing Group 2007).
[84] Nassim Nicholas Taleb, 'Black Swans and the Domains of Statistics' (2007) 61 The American Statistician 198.
[85] Paul Goodwin and George Wright, 'The Limits of Forecasting Methods in Anticipating Rare Events' (2010) 77 Technological Forecasting and Social Change 355.
[86] NN Taleb, *Antifragile: Things That Gain from Disorder* (Random House Publishing Group 2012).

irreversible and where cost-benefit analysis no longer applies in the face of what are effectively infinite costs, such as the loss of animal (or the human) species.

V.V. CASE STUDY: EXISTENTIAL RISKS

Within long-term and/or low-probability event forecasts, there exists a specific example that has additional specific associated complexities and has recently been heavily publicized: existential risks. These are defined as catastrophic events threatening a significant portion or all of human life, or its societal structure. As we will shortly discuss, existential risks span a wide array of dangers, from harmful artificial intelligence to nuclear weapons and global epidemics. Anticipating such events directly influences the prospects of humankind, as the realization of even a single event has extremely long-lasting and far-ranging effects on the course of civilisation. In this section, we discuss how, while predictions regarding existential risks are the most extreme area of forecasting, many of the concepts introduced so far such as biases still apply; the following section will use concrete examples to illustrate the diffulties and solutions in forecasting such events.

Beyond philosophical considerations, attempting to predict such catastrophic events has very practical implications. Because they are typically black swan events, it is important not only to mitigate existential risks, but also to steer research studies and public awareness and conversation towards resilience. Contemporary examples of such preventive planning include the precautionary principle widely applied in the European Union[87], or the German civil defence plan suggesting that each citizen should have enough water and food to survive for a week without assistance[88].

Existential risks are particularly complicated to predict, mainly because they compound all of the difficulties discussed previously, including those of long-term and black swan forecasting. By definition, empirical evidence relating to existential risks is extremely sparse. Typically, the only response that can currently be proposed is to apply methods mentioned above with extreme scrutiny. This could imply making exhaustive lists of milestones, or combining all

[87] Treaty on the Functioning of the European Union, Part Three, Title XX, Article 191.

[88] Caroline Copley, 'Germany to Tell People to Stockpile Food and Water in Case of Attacks: FAS' *Reuters* (2016).

possible methods of scenario generation rather than relying on a single method that may be optimal under different circumstances[89].

A significant part of the interest in existential risk forecasting lies in the specificities of tackling such questions. First of all, fear of non-survival leads to irrational behaviour: information elicitation methods may become inherently biased because participants feel that they increase their chances of survival by choosing actions that would be irrational under any other circumstance. For instance, it would be impossible to implement a prediction market to answer questions concerning catastrophic risks: the value, and perhaps even the meaning, of a monetary payoff may be hard to define. An alternative proposal could then be to replace the monetary payoff with access to special catastrophe-proof bunkers. However, in this case, participants would have an interest in trying to manipulate the market, on the off-chance that receiving such a ticket may be a survival necessity.

Secondly, thresholds of no-return imply that most catastrophic events have irreversible effects. Contrary to elections that are organized regularly every few years, or to stock prices that may rebound to some value after a small or even large crash, existential risks by their very nature imply non-repeatability. For example, a large variety of catastrophes studied result in the death of the entire human population, such that the species does not get the opportunity to even attempt reconstruction. In other cases, feedback loops imply a no-return situation from a single lower-impact event. An early example of this is the theory of Mutually Assured Destruction developed during the Cold War, when the USA or Soviet Union could potentially retaliate to a nuclear aggression by using their own nuclear arsenal, initiating a spiral until the entire nuclear arsenals of both superpowers were engaged.

Finally, on a more general note, some existential risks are intrinsically single events, such that their consequences are always at least partly speculative: there may be no milestone to help forecasting or historical data significant enough. It would be extremely difficult to predict the effects of a kilometre-radius asteroid impact, for instance, in today's world.

[89] Thomas Rowe and Simon Beard, 'Probabilities, Methodologies and the Evidence Base in Existential Risk Assessments'.

V.VI. EXAMPLES OF EXISTENTIAL RISKS

This section presents several areas of particular interest for forecasting existential risks; these examples are chosen both for the breadth of existing research and for the wide-ranging exposure they have had in the media, making their forecasting particularly important to the wider public. While the field of existential risks is much larger, we endeavour to show with these examples how predictions regarding such events fit within our analysis of forecasting techniques.

i. Artificial Intelligence

The current hot-topic in existential risks is that related to artificial intelligence, for instance from autonomous weapons[90] to artificial general intelligence[91]. Many different approaches have been suggested for forecasting the development of artificial intelligence and the occurrence of existential catastrophes. For example, various methods have been published to model pathways, i.e. map logical connections between the different required milestones, as in Figure 5[92]. These can be applied to specific topics, such as pathways to failure of containment, the set of measures restricting an AI system's reach into external IT systems.

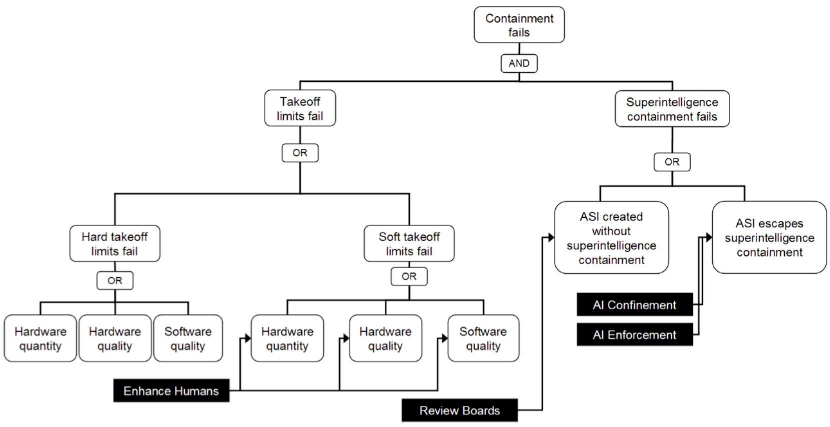

Figure 5: Model pathways towards containment failure for artificial intelligence; reproduced from Barrette & Baum (2017).

With an important scientific community now studying the topic, opinion surveys are becoming more frequent and detailed. These can be used to assign probabilities to developments and

[90] Future of Life Institute, 'An Open Letter to the United Nations Convention on Certain Conventional Weapons' (2017).
[91] Tom Chatfield, 'How Much Should We Fear the Rise of Artificial Intelligence?' *The Guardian* (2016).
[92] Anthony M Barrett and Seth D Baum, 'A Model of Pathways to Artificial Superintelligence Catastrophe for Risk and Decision Analysis' (2017) 29 Journal of Experimental & Theoretical Artificial Intelligence 397.

milestones, such as those formulated in pathway models. For instance, AI experts have been polled about the "earliest [date] that machines will be able to simulate learning and every other aspect of human intelligence" or the "earliest [date] we will understand the architecture of the brain sufficiently to create machine simulation of human thought"[93].

Mathematical and simulation tools have also been used to model conditions under which an AI arms race is likely to occur. For instance, Armstrong, Bostrom and Shulman counter-intuitively found that competition between teams developing AI is more dangerous if information about progress is available to all teams[94]. The use of a toy-model that is arguably simplistic is however an important limitation in this context.

This multitude of approaches and the number of studies undertaken on AI safety since the 1950s also means that we can now assess the forecasts made over the years. In particular, Armstrong, Sotala and ÓhÉigeartaigh have studied multiple types of forecasts about AI, including causal and non-causal models, philosophical arguments, etc. Their examination of expert and non-expert judgment has revealed that both exhibit a propensity to predict the development of human-level AI 15 to 25 years in the future, as can be seen in the figure 6. However, these predictions were found not to follow the Maes-Garreau law, a bias for predicting that a technology such as AI will reach maturity in time to save the predictors from death[95]. We conclude that expert surveys likely have little value in the context of AGI forecasting, although the precise reason for inaccuracy is not fully understood.

---

[93] Vincent C Müller and Nick Bostrom, 'Future Progress in Artificial Intelligence' (2014) 1 AI Matters 9.
[94] Stuart Armstrong, Nick Bostrom and Carl Shulman, 'Racing to the Precipice: A Model of Artificial Intelligence Development' (2016) 31 AI and Society 201.
[95] Stuart Armstrong, Kaj Sotala and Seán S Ó hÉigeartaigh, 'The Errors, Insights and Lessons of Famous AI Predictions – and What They Mean for the Future' (2014) 26 Journal of Experimental & Theoretical Artificial Intelligence 317.

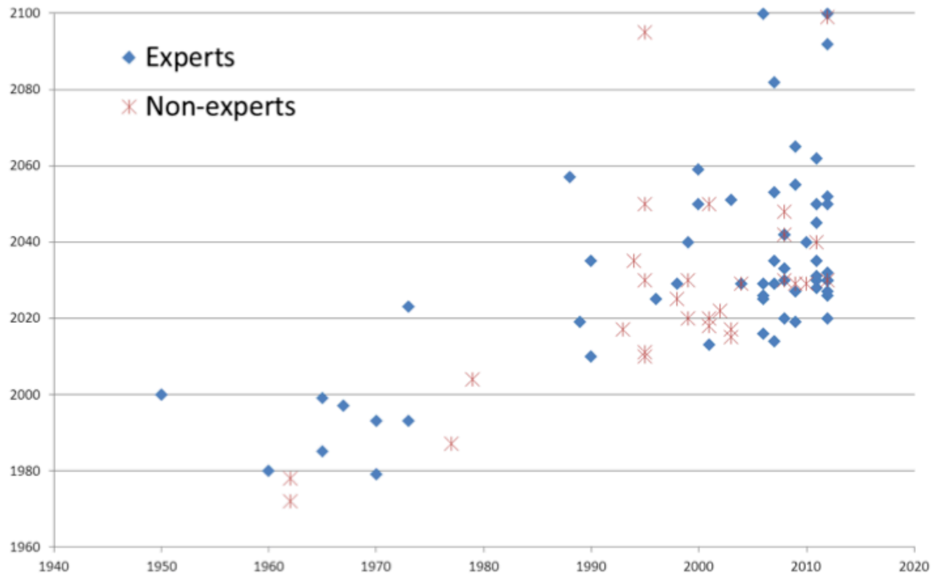

Figure 6: Estimated year for the emergence of human-level artificial intelligence, against date of prediction; reproduced from Armstrong, Sotala & ÓhÉigeartaigh (2014)

ii. Nuclear Catastrophes

Historically, a prime example of existential risks has been those linked to nuclear catastrophes. For instance, U.S. President John F. Kennedy estimated the probability of a nuclear holocaust by stating "somewhere between one out of three and even"[96]. Beyond the lack of time-scale on this prediction, this example shows the importance of the perceived risk.

One of the reasons behind this particular attention is that humanity has already passed many of the key milestones towards catastrophic nuclear explosions. It has developed atomic bombs thousands of times more destructive than so-called conventional weapons, and bombs thousands of times more powerful than the former. It has used such weapons on civilian populations in the bombings of Hiroshima and Nagasaki. It has developed variants of such weapons to create the Nuclear Triad, such that any region of Earth can be targeted at any time. Finally, it has stockpiled these weapons in quantities sufficient for catastrophic results to happen[97].

[96] Future of Life Institute (n 90).

[97] N Beckstead and others, 'Unprecedented Technological Risks' (2014).

On the other hand, the rarity and controllability of resources, such as Uranium-235 and Plutonium-239, as well as political will and stringent safety policies have led to a decrease in perceived risk and to the avoidance of such catastrophe since the Trinity Test in 1945[98].

iii. Climate Change

A final example of existential risk is that of catastrophic climate change. This is currently one of the most widely studied and publicized topics: significant resources are being allocated to anticipate and mitigate the risks linked to the phenomenon of global warming, such that the accuracy of the forecasts directly influences the effectiveness of prevention measures and environmental planning policies. It is also a typical case of ongoing, continuous phenomenon, where new data is continuously available: forecasts can be assessed and updated virtually in real time. In particular, multiple recent studies have focused on evaluating previous predictions. For instance, Fildes and Kourentzes have shown that so-called time-series models produce high-accuracy forecasts, and suggest that merging them with atmospheric-ocean general circulation models could improve this further[99]. Another study examined the accuracy of early climate forecasts, which are often criticized as overestimates of the effects of climate change: comparing forecasts dating from 1996 to the data acquired until 2012, the accuracy of the predictions was shown to outperform simple interpretation of today's climate model. The same study suggested ways of combining modelling with up-to-date experimental data[100]. These examples underscore the importance of applying a large variety of forecasting methods to tackle many aspects of an issue, and to do so repeatedly as the issue progresses and new information becomes available.

[98] Richard Ned Lebow, 'Is Crisis Management Always Possible?' (1987) 102 Political Science Quarterly 181; Richard Ned Lebow, *Avoiding War, Making Peace* (Springer International Publishing 2018).
[99] Robert Fildes and Nikolaos Kourentzes, 'Validation and Forecasting Accuracy in Models of Climate Change' (2011) 27 International Journal of Forecasting 968.
[100] Myles R Allen, John FB Mitchell and Peter A Stott, 'Test of a Decadal Climate Forecast' (2013) 6 Nature Geoscience 243.

VI. POLICY RECOMMENDATIONS

VI.I. POLICY GOALS

We now turn to answering some of the issues raised in this paper, in the form of goals for future policy proposals, as well as a set of policy recommendations to fulfill these goals. At this point, it is necessary to restate the aim of this policy paper. The starting point of our discussion was the realization that we, as individuals and as groups, are rather inaccurate forecasters and have only started assessing the precision and quality of our predictions. Recent studies have established that there are numerous ways in which significant improvements can be made.

It is also important to reiterate the importance of proposing policy recommendations. While recent academic work needs to be refined and extended in coming years, the main motivation for such academic research has always been the practical impact of potential improvements made to forecasting techniques. Existing research already produces some clear directions for improving real-world predictions. Pure research and raising awareness for its results should therefore be confirmed by implementing targeted policy to achieve the improvements previously discussed.

In this context, we here present a number of goals for policy proposals. In the subsequent section, we will specify detailed examples of recommendations for each objective. These goals are as follows:

1) **Incentivize and raise awareness for appropriate forecasting methodology**. The fundamental obstacle to more accurate predictions is the lack of understanding of basic forecasting concepts and requirements, by the public and professionals alike. Immediate improvements, encouraged first by the leadership of corporations and public agencies, could include avoiding vague phrasing, stating precise timescales and validation methods explicitly.

2) **Create tools and frameworks that allow people to avoid their individual and collective biases.** These tools could be used to alleviate individual and group-level bias, but would also allow individuals to be aware of other people's biases. It could also allow forecasters to ask more specific and falsifiable questions, that are easy to interpret by others, as well as to overcome frame blindness.

3) **Implement systematic checks and feedback for all predictions made within organizations.** A crucial milestone in the improvement of forecasting would be the generalized acceptance of values such as transparency and continuous assessment. Historically, lack of transparency has been the main challenge to the implementation of any worthy prediction system within public institutions and private companies.

4) **Subsidize research into the working conditions of various forecasting techniques, and in particular prediction markets.** A major obstacle to the practical use of prediction methods is the lack of understanding of the environments in which they are accurate. For instance, for prediction markets, research could tackle a wide range of conditions, including market liquidity, time ranges, information availability, or participant selection, in order to evaluate the parameter space in which prediction markets can be used reliably.

5) **Incentivize the public release of prediction data and forecasts from private companies.** A huge accelerator for improvements in forecasting techniques would be to adopt an open-source approach, to both the methods used and the data collected. Such a paradigm change, pushed by governmental and supranational entities for instance, would allow a faster aggregation of the results of individual studies, and potentially spawn a new industry linked to services in forecasting.

VI.II. SPECIFIC POLICY RECOMMENDATIONS

In order to fulfil these goals and achieve the improvement of forecasting accuracy and applications, specific policies will have to be implemented. These may apply to a wide variety of contexts, including public and private institutions, individuals in personal and professional settings, etc. Beyond these general goals, however, specific policies are also required to fuel a healthy debate about the role of forecasting, illustrate what actions could be taken to improve our capabilities, and prove the potential of enhanced prediction methods and skills.

To incentivize and raise awareness for appropriate forecasting methodology, we propose to:

- **Introduce basic training in forecasting methods**, including Bayesian statistics and probability. Although single-intervention training has been shown to be successful, other studies recommend prolonged training. This could include: (i) adding/modifying school curricula to include examples of problems solved using Bayesian inference, including specifically referring to forecasting, (ii) including these within university syllabi for topics related to forecasting (e.g. management studies, law, etc.), and (iii) offering such training as part of professional training and continuing education. Both governments and business executives can set up and promote these actions. Training should go beyond basic Bayesian statistics into complementary topics, including those pertaining to forecasting in situations with limited information.

- **Fund IARPA-style studies within large governmental agencies.** In the UK, forecasting assessment has been demonstrated in the NHS for predicting the availability of future medical technologies; extending this to broader forecasting targets, and beyond the NHS, should be straightforward using the scalable methods discussed. Open-sourcing the results would demonstrate the value of increased transparency and of assessing forecasts, leading to a culture change even beyond the target agencies.

- **Create open working cultures.** The main goal is to avoid using vagueness as way to hedge risks or to manage expectations. For instance, independent referees could be named to scour for ambiguous statements in guidance and forecasting documents. Rewards could be introduced within governmental bodies and inside private companies for attempting difficult predictions and for particularly transparent forecasting. These could be monetary at first, and later become part of criteria for professional progression.

To alleviate or reduce the influence of biases on forecasting, we propose the following:

- **Use a tool that can reduce groupthink and group bias in organisations that are making forecasts**. An example of a tool that would reduce the biases and improve forecasting accuracy is the previously discussed "**Dot Collector**"[101]. This tool harnesses the power of algorithms, honesty, meritocracy and democracy to alleviate both group bias and individual bias. Ray Dalio has already announced that he plans to release this software soon. We recommend that governments and companies should implement practices such as using the "Dot Collector" or similar, manually made software, that can improve forecasting accuracy. Such policies will make for better organizational predictions.

- Forecasters should use a **list of cognitive biases** to avoid their own biases, but also to take into account other people's biases when making predictions. Specifically, forecasters can

[101] Dalio (n 38).

make or use the list in **Appendix 1** i.e., **The Bias List for Mindful Forecasting** (BLMF), in two main ways. First, they can make sure that they are not succumbing to any of these biases while making their prediction. Second, and more importantly, if they are predicting the decisions or behaviours of other people, they can take into account that these decisions and behaviours are largely the product of other people's biases. Altogether this would make for more accurate forecasts. Companies and governments should implement policies where using the BLMS is stated as a preferable or mandatory approach.

- Forecasters should **ask precise questions**, to avoid people's natural tendency to choose simpler loosely-defined questions rather than clear-cut ones[102]. First, forecasters have to ask questions that are **falsifiable.** This means **precisely setting out the parameters that would make a forecast wrong or right**. Therefore asking "Is America going to attack Iceland?" is not precise, because it gives us no indication of why, when, or where this would happen. However, asking the question "Is the USA going to attack Iceland in January 2030?" provides clear definitions for what is meant by America, and when America may attack Iceland. Thus companies and governments should implement policies which make it mandatory to make predictions falsifiable in order to be considered a legitimate forecast, in the same way it is mandatory to make research falsifiable for it to be considered as science. Again, independent or third-party referees, or even AI systems in the near future, could easily flag statements which are ambiguous and cannot be falsified.

- Another effective way of asking more precise questions is by having **operational definitions** for every term used in forecasting questions[103]. Operational definitions should include: 1) the **validation test** needed to determine the nature of a concept; 2) a **concept's properties,** e.g. quantity, duration, size, etc. If a forecaster was to ask "Is America going to to war with Iceland in the next 10 days?", they would need to define all the concepts in this question. To give an example, the operationalization of war could be America officially declaring war on Iceland. It is important because one concept can have multiple operational definitions in different contexts. Thus, governments and companies should implement policies which make it mandatory for predictions to have operational definitions for every term that is used in the forecast.

---

[102] Tetlock and Gardner (n 2).
[103] SS Stevens, 'The Operational Definition of Psychological Concepts.' (1935) 42 Psychological Review 517.

- Finally, particular attention should be placed in **overcoming frame blindness**, and thinking through the multiple causal pathways by which events, and particularly rare events (including existential risks), could occur. Creative approaches, such as performing binary predictions with varied phrasings across multiple platforms, would help in identifying the pathways that lead to highly unlikely events, as well as milestones for such pathways.

We also recommend the implementation of systematic checks of all predictions made within organizations:

- The British government should **carry out a large-scale evaluation of existing forecasting methods**, similar to IARPA's work. Some governmental bodies are particularly well-suited, including the NHS and intelligence agencies such as GCHQ. Both rely heavily on forecasting, respectively in medical technology and geopolitics, and would provide high-impact and heavily publicized illustrations of the concepts discussed here.

- **Systematically assess the efficiency of debiasing training** and other forecasting improvement methods. Any organization providing debiasing training to their staff should simultaneously assess its impact. The simplest tool would be to provide basic guidelines and expert help as a public service. For-profit consulting firms may also develop around the business of providing such debiasing training and impact evaluation.

- As an illustration of this, conduct a **large-scale review of all guidance for civil servants** concerning forecasting and subjective judgements, experiment with specialized training, and systematically assess its impact.

- More generally, a wide **study of what biases affect each information elicitation/aggregation method** would be helpful to define the ideal technique for a range forecasting purposes. This would radically decrease the complexity of choosing and applying forecasting improvements for non-experts.

- **Create a peer review/fact-checking system for forecasters.** The ideal scenario for researchers in the field would be a method for rapidly verifying their past forecasts, and rate them against their peers' predictions. While there may be some cultural barrier, providing a unified tool, such as an online comparison database, would remove any technical difficulty.

There are several policy changes pertaining specifically to prediction markets that would substantially help with the long-run improvement of forecasting techniques. In this area, we recommend:

- **Strengthening legislative support for prediction markets, particularly in the United States.** Ideally, the US government should create a specific legislative framework that allows for legal prediction markets while maintaining the desired limitations on gambling. In the absence of legislation, progress could still be made by expanding and systematising the use of "no-action letters" from the CFTC, which exempt a particular market from regulatory oversight. The United Kingdom has unusually lenient regulations on gambling, so legislative change may not be necessary here.
- **Using prediction markets to inform economic forecasts in central banks and Treasuries.** Market-derived predictions are already in use to some extent – the Bank of England, for instance, conditions its economic forecasts on the path for Bank Rate implied by yields in the UK bond market. Making more explicit use of market forecasts could be a useful and politically feasible way to extend the use of more systematic forecasting techniques, such as prediction markets, in government. These institutions could also consider setting up dedicated prediction markets, rather than using implied predictions from already-traded assets, where this would improve forecasts.
- **Using effective forecasting methods, including prediction markets, in firms.** This could be done in particular through the work of the Productivity Leadership Group, which is a government- supported initiative to improve productivity in Britain's businesses. It gives various recommendations and training to managers and owners of companies. Improving the accuracy of forecasting could improve the productivity of businesses, particularly those that require long-term decision-making, and so the group should consider including information about forecasting methods in its corporate training.
- **In general, when prediction markets are used, framing them explicitly as tools to improve predictions rather than ways to make money off predictions.** The Policy Analysis Market failed largely because it was perceived as unjust to profit off predictions about events like terrorism, even though accurate predictions from the market could have helped to prevent terrorist incidents. To avoid PR backlash, organisations using prediction markets -

particularly if they deal with sensitive topics - should emphasise the social or private benefits from more accurate forecasting, and the fact that the element of profit is simply a tool to make predictions more accurate.

- **Introducing internal, play-money prediction markets or prediction polls in organisations such as the civil service and private firms.** These types of prediction methods entail the least risk for organisations, and as discussed are still able to capture some of the benefits of larger prediction markets. Use of these types of prediction markets could act as a test case and a basis for research into the effectiveness of different prediction methods in practical settings.
- **Funding and incentivising research into practical applications of prediction markets.** Theoretical research and lab experiments on prediction markets have certainly been useful, but at present evidence on their performance in different real-world settings is somewhat more scarce. Funding should particularly be oriented towards research where private companies agree to test systematic prediction methods.
- Finally, to overcome the weaknesses of prediction markets, **funding of complementary prediction methods** is necessary. Many situations where information is not widely distributed require further emphasis on information elicitation, overcoming frame blindess, and choosing the appropriate forecasters, whether among experts or situation-specific participants. Because of their mostly experimental nature, such attempts are best suited in the context of government-funded studies.

As we have mentioned previously, research into forecasting techniques should be strongly encouraged, in both academic and corporate settings. The current body of work on the topic is quite recent but has already yielded important and conclusive results; these should be used as proof of the enormous potential of this emerging field.

More generally, we strongly advise any research or implementation of forecasting methods to be realized under a strict scientific approach, with a fully non-partisan agenda. The goal here is to improve the tools available to forecasters and the value added by predictions through higher accuracy and reliability, rather than to create a method for confirming personal beliefs and objectives.

VI.III. PREDICTION VERSUS ACTING

Although an aside, it is important to differentiate between forecasting, decision-making and acting. Forecasting corresponds specifically to prediction attempts, to finding a measure of probability for a given event. Although a loose definition of decision-making may include forecasting, we can define decision-making more strictly as setting a threshold for actions to be implemented, in response to predicted events. Acting then refers to setting up a system that responds to the forecasts in a way that was decided previously. Each requires a specific set of skills and cognitive aptitudes, although there are of course significant overlaps. Here, we wish to stress that their goals and the approaches to their study are distinct.

This policy paper focuses voluntarily on forecasting only, and dismisses any considerations linked to pure decision-making. Forecasting is already a huge subject in its own right, with enormous philosophical and practical implications. Moreover, it is a necessary input to decision-making, and in many ways is more fundamental than the latter. In that sense, forecasting offers the most obvious potential for improvement. The field of decision-making is perhaps also currently more widely known, through many studies and guides in the context of leadership.

There is however obviously a feedback loop between prediction and decision-making: there exist many situations in which predictions depend on decisions, and vice-versa. For instance, a company may set up a prediction market to help forecast future sales, and then decide that some action is required if the expected numbers are too low, which will lead to an immediate response and adjustment by the prediction market to update for this new information.

On the other hand, in practice, forecasting and decision-making roles are typically well-defined and split between different groups. For instance, research institutes make predictions for government agencies, which then debate and decide on the action to take. This splitting has the significant advantage of making sure that the elements in this feedback loop are independent, implying that they can somewhat check each other. An important disadvantage is the requirement for good communication and understanding between these groups to ensure high-quality feedback. It is particularly important that decision-makers and actors truly understand the meaning of predictions. The usefulness of forecasting also heavily relies on predictions being communicated to the relevant decision-makers at the relevant times, in a suitable format.

More generally, decision-makers may require additional training or tools to understand the full implications of forecasts being brought to their attention.

Our hope is then that better predictions will lead to better decisions, and that the content presented in this policy paper can be linked to debates on decision-making. We should also note that making verifiable and reliable predictions is an important task for the future: forecasting based on scientific methods and backed by objective results from research will lead in more trust in predictions, and such trust would be hard-earned rather than blind.

VII. CONCLUSION

VII.I. PAPER SUMMARY

Our approach and logical developments are based on the necessity to view forecasting as a truly scientific field. The main premise for this policy paper is that we, as individuals and as a society, are rather inaccurate forecasters. We have therefore demonstrated that there are clear paths for improvement, based on solid scientific studies, and we have discussed how any improvement has direct positive effect on individuals and groups.

With this in mind, the first objective of this policy paper has been to explain how forecasting is embedded in the daily activities of individuals and organizations alike. Then, arguing the importance and possibility of improving prediction methodology, we have summarized the current state-of-the-art in academic research and applications. We have used a first-principles approach, starting from the fundamental cognitive processes and biases and building up towards the accuracy of complex prediction methods. We have assessed the particular mechanism of prediction markets; these generally perform better than alternative methods of group forecasting, both in theory and in practice, although other crowd-based prediction methods such as prediction polls are also comparably effective. A caveat to this success is the presence of unique difficulties with long-term forecasting, which is much less amenable to the use of prediction markets and similar methods; we suggest some alternative approaches, such as scenario generation and debiasing training.

Finally, we wish to summarize the goals set in this paper for future policies. The most basic is to incentivize and raise awareness for proper forecasting methodology: the biggest current obstacles are the lack of acknowledgement of and understanding that subjective judgements can and should be improved. In keeping with the first-principles approach, we recommend building a framework for the analysis of forecasting methods in terms of individual and collective biases. More generally, an important amount of work has to be done to increase transparency in all organizations, including the implementation of systematic checks for predictions made. Beyond solely implementing protocols, it requires a true change in organizations' culture towards verifiability and transparency. A possible step would be to incentivize the public release of prediction data and forecasts, including from private companies, to assess their accuracy and encourage a wider reflection about the impact of forecasting. In order to fully convince citizens of the value of good forecasting, governments may wish to allocate significant resources to subsidize research into practical implementations of forecasting methods, including prediction markets.

VII.II. AFTERWORD: IMPACT ON FUTURE GENERATIONS

Improving forecasting techniques is particularly important today: as presented throughout this policy paper, advancements in prediction methods are among the topics with the largest potential to change the course of society in coming years. Enabling better judgement has an enormous impact on all levels of society, and in all circumstances: a family making sounder financial forecasts, a factory predicting raw material requirements more accurately, voters analyzing politicians' agendas more objectively, a judge predicting successful reinsertion probabilities more accurately, etc. Each of these leads to a tangible impact, for instance by avoiding missed mortgage payments, freeing up cash for additional equipment purchases, voting for a candidate whose policy proposals are closer to one's priorities, or giving a prisoner an earlier second chance.

Thinking of forecasting as a scientific tool also implies a strong cultural change, towards more verifiability and transparency, and would lead us to view society and causal relationships within it more accurately. Improving forecasting therefore means that positive feedback effects can be strengthened: improved forecasting leading to refined decisions, which can then be re-evaluated through further subjective judgements, such that the improvement in the methods used increases the value added with each cycle.

We can therefore view improvements in forecasting techniques not as a one-time increase, but rather as a steepening of the slope of progress. In this sense, it is important to pursue this goal not only for the short-term impact, but more significantly for the value of improving the tools that will be used by future generations (a topic of particular interest to the Wilberforce Society[104]). It is therefore up to our current generations to prove and demonstrate practically that research in forecasting can yield significant and tangible positive benefits. The use and applications of advanced forecasting techniques must consequently go much beyond academic research. In the medium-term, this would culminate in high-profile successes to ensure that this

[104] Natalie Jones, Mark O'Brien and Thomas Ryan, 'Representation of Future Generations in United Kingdom Policy-Making' (2018) 102 Futures 153.

approach to forecasting quickly permeates public understanding and becomes natural for each of us.

BIBLIOGRAPHY

Allen MR, Mitchell JFB and Stott PA, 'Test of a Decadal Climate Forecast' (2013) 6 Nature Geoscience 243

Amer M, Daim TU and Jetter A, 'A Review of Scenario Planning' (2013) 46 Futures 23

Ang A, Piazzesi M and Wei M, 'What Does the Yield Curve Tell Us about GDP Growth?' (2006) 131 Journal of Econometrics 359

Armstrong S, Bostrom N and Shulman C, 'Racing to the Precipice: A Model of Artificial Intelligence Development' (2016) 31 AI and Society 201

Armstrong S, Sotala K and Ó hÉigeartaigh SS, 'The Errors, Insights and Lessons of Famous AI Predictions – and What They Mean for the Future' (2014) 26 Journal of Experimental & Theoretical Artificial Intelligence 317

Atanasov P and others, 'Distilling the Wisdom of Crowds: Prediction Markets vs. Prediction Polls' (2017) 63 Management Science 691

Bar-Haim Y and others, 'Threat Related Attentional Bias in Anxious and Non Anxious Individuals: A Meta-Analytic Study' (2006) 133 Psychological Bulletin, 133, 1 - 24 (2007)

Barrett AM and Baum SD, 'A Model of Pathways to Artificial Superintelligence Catastrophe for Risk and Decision Analysis' (2017) 29 Journal of Experimental & Theoretical Artificial Intelligence 397

Barron G and Leider S, 'The Role of Experience in the Gambler's Fallacy' (2010) 23 Journal of Behavioral Decision Making 117

BBC News Americas, 'Pentagon Axes Online Terror Bets' *BBC News* (29 July 2003)

Beckstead N and others, 'Unprecedented Technological Risks' (2014)

Beebee H, Hitchcock C and Menzies P (eds), *The Oxford Handbook of Causation*, vol 1 (Oxford University Press 2010)

Berg JE and Rietz TA, 'Prediction Markets as Decision Support Systems' (2003) 5 Information Systems Frontiers 79

Blackmond Laskey K and others, 'Graphical Model Market Maker for Combinatorial Prediction Markets' (2018) 63 Journal of Artificial Intelligence Research 421

Borger J, 'All Bets Are off on Terror, Rules Pentagon' *The Guardian* (30 July 2003) <https://www.theguardian.com/world/2003/jul/30/usa.julianborger>

Campbell JE and Garand JC, *Before the Vote: Forecasting American National Elections* (Sage Publications 2000)

Chang W and others, 'Developing Expert Political Judgment: The Impact of Training and Practice on Judgmental Accuracy in Geopolitical Forecasting Tournaments' (2016) 11 Judgment and Decision Making 509

Chatfield T, 'How Much Should We Fear the Rise of Artificial Intelligence?' *The Guardian* (2016)

Chernoff H, 'A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the Sum of Observations' (1952) 23 The Annals of Mathematical Statistics 493

Copley C, 'Germany to Tell People to Stockpile Food and Water in Case of Attacks: FAS' *Reuters* (2016)

Cowgill B, Wolfers J and Zitzewitz E, 'Using Prediction Markets to Track Information Flows: Evidence from Google' (2009)

Dalio R, 'How to Build a Company Where the Best Ideas Win' (2017)

——, *Principles: Life and Work* (Simon & Schuster 2017)

Deck C, Lin S and Porter D, 'Affecting Policy by Manipulating Prediction Markets: Experimental Evidence' (2013) 85 Journal of Economic Behavior & Organization 48

Etchells P, 'Declinism: Is the World Actually Getting Worse?' *The Guardian* (16 January 2015) <https://www.theguardian.com/science/head-quarters/2015/jan/16/declinism-is-the-world-actually-getting-worse>

Fama EF, 'Efficient Capital Markets: A Review of Theory and Empirical Work' (1970) 25 The Journal of Finance 383

Fildes R and Kourentzes N, 'Validation and Forecasting Accuracy in Models of Climate Change' (2011) 27 International Journal of Forecasting 968

Firat AK, Madnick S and Woon WL, 'Technological Forecasting–A Review' (2008)

Frisch D and Baron J, 'Ambiguity and Rationality' (1988) 1 Journal of Behavioral Decision Making 149

Furnham A and Boo HC, 'A Literature Review of the Anchoring Effect' (2011) 40 The Journal of Socio-Economics 35

Future of Life Institute, 'An Open Letter to the United Nations Convention on Certain Conventional Weapons' (2017)

Gneiting T and Raftery AE, 'Strictly Proper Scoring Rules, Prediction, and Estimation' (2007) 102 Journal of the American Statistical Association 359

Goodwin P and Wright G, 'The Limits of Forecasting Methods in Anticipating Rare Events' (2010) 77 Technological Forecasting and Social Change 355

Hanson R, 'Decision Markets' (1999) 14 IEEE Intelligent Systems and their Applications 16

——, 'Combinatorial Information Market Design' (2003) 5 Information Systems Frontiers 107

——, 'The Policy Analysis Market (A Thwarted Experiment in the Use of Prediction Markets for Public Policy)' (2007) 2 Innovations: Technology, Governance, Globalization 73

——, 'Shall We Vote on Values, But Bet on Beliefs?' (2013) 21 Journal of Political Philosophy 151

Hanson R, *The Age of Em: Work, Love, and Life When Robots Rule the Earth* (Oxford University Press 2016)

Hayek FA, 'The Use of Knowledge in Society' (1945) 35 The American Economic Review 519

Institute For The Future, 'The American Future Gap Survey' (2017)

Jacobs V, 'Prediction Markets: How They Work and How Well They Work' (Catholic University of Louvain 2009)

Jones N, O'Brien M and Ryan T, 'Representation of Future Generations in United Kingdom Policy-Making' (2018) 102 Futures 153

Kahneman D, *Thinking, Fast and Slow* (Farrar, Straus and Giroux 2011)

Kahneman D and Tversky A, 'Subjective Probability: A Judgment of Representativeness' (1972) 3 Cognitive Psychology 430

——, 'On the Psychology of Prediction.' (1973) 80 Psychological Review 237

——, 'Prospect Theory: An Analysis of Decision under Risk' (1979) 47 Econometrica 263

——, 'Choices, Values, and Frames.' (1984) 39 American Psychologist 341

Lebow RN, 'Is Crisis Management Always Possible?' (1987) 102 Political Science Quarterly 181

——, *Avoiding War, Making Peace* (Springer International Publishing 2018)

Maloney MT and Mulherin JH, 'The Complexity of Price Discovery in an Efficient Market: The Stock Market Reaction to the Challenger Crash' (2003) 9 Journal of Corporate Finance 453

Manski C, 'Interpreting the Predictions of Prediction Markets' (2004)

Müller VC and Bostrom N, 'Future Progress in Artificial Intelligence' (2014) 1 AI Matters 9

Page L and Clemen RT, 'Do Prediction Markets Produce Well-Calibrated Probability Forecasts?' (2013) 123 The Economic Journal 491

Pearl J, *Causality*, vol 19 (Cambridge University Press 2009)

Plott C and Chen K-Y, 'Information Aggregation Mechanisms: Concept, Design and Implementation for a Sales Forecasting Problem'

Plous S, *The Psychology of Judgment and Decision Making.* (Mcgraw-Hill Book Company 1993)

Poon C, Koehler D and Buehler R, 'On the Psychology of Self-Prediction: Consideration of Situational Barriers to Intended Actions' (2014) 9 Judgment and Decision Making 207

Porter AL and others, 'Technology Futures Analysis: Toward Integration of the Field and New Methods' (2004) 71 Technological Forecasting and Social Change 287

Rhode P and Strumpf K, 'Historical Political Futures Markets: An International Perspective' (2008)

Rhode PW and Strumpf KS, 'Historical Presidential Betting Markets' (2004) 18 The Journal of Economic Perspectives 127

Rothschild D, 'Forecasting Elections' (2009) 73 Public Opinion Quarterly 895

Rowe T and Beard S, 'Probabilities, Methodologies and the Evidence Base in Existential Risk Assessments'

Salo A, Gustafsson T and Ramanathan R, 'Multicriteria Methods for Technology Foresight' (2003) 22 Journal of Forecasting 235

Sandberg A and Bostrom N, 'Whole Brain Emulation: A Roadmap' (2008)

Satopää VA and others, ‘Probability Aggregation in Time-Series: Dynamic Hierarchical Modeling of Sparse Expert Beliefs’ (2014) 8 The Annals of Applied Statistics 1256

Servan-Schreiber E and others, ‘Prediction Markets: Does Money Matter?’ (2004) 14 Electronic Markets 243

Silverman C, ‘The Backfire Effect’ (*Columbia Journalism Review*, 2011) <https://archives.cjr.org/behind_the_news/the_backfire_effect.php>

Snowberg E, Wolfers J and Zitzewitz E, ‘Prediction Markets for Economic Forecasting’, *Handbook of Economic Forecasting*, vol 2 (2013)

Stevens SS, ‘The Operational Definition of Psychological Concepts.’ (1935) 42 Psychological Review 517

Sunstein CR and Kuran T, ‘Availability Cascades and Risk Regulations’ (1999) 51 Stanford Law Review 683

Surowiecki J, *The Wisdom of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations.* (Doubleday & Co 2004)

Szabo N, ‘Falsifiable Design: A Methodology for Evaluating Theoretical Technologies’ (2007) <http://unenumerated.blogspot.co.uk/2007/02/falsifiable-design-methodology-for.html>

Taleb NN, *The Black Swan: The Impact of the Highly Improbable* (Random House Publishing Group 2007)

Taleb NN, ‘Black Swans and the Domains of Statistics’ (2007) 61 The American Statistician 198

Taleb NN, *Antifragile: Things That Gain from Disorder* (Random House Publishing Group 2012)

Taleb NN and Tetlock PE, ‘On the Difference between Binary Prediction and True Exposure with Implications for Forecasting Tournaments and Decision Making Research’ [2013] SSRN Electronic Journal

Tetlock PE, *Expert Political Judgment* (STU-Stud, Princeton University Press 2005)

Tetlock PE and Gardner D, *Superforecasting: The Art and Science of Prediction.* (Crown Publishers/Random House 2015)

Todd P and others, *Simple Heuristics That Make Us Smart* (1999)

Turner ME and Pratkanis AR, 'A Social Identity Maintenance Model of Groupthink' (1998) 73 Organizational Behavior and Human Decision Processes 210

——, 'Twenty-Five Years of Groupthink Theory and Research: Lessons from the Evaluation of a Theory' (1998) 73 Organizational Behavior and Human Decision Processes 105

Tversky A and Kahneman D, 'Availability: A Heuristic for Judging Frequency and Probability' (1973) 5 Cognitive Psychology 207

Tversky A and Kahneman D, 'Judgment under Uncertainty : Heuristics and Biases Linked References Are Available on JSTOR for This Article : Judgment under Uncertainty : Heuristics and Biases' (1974) 185 Science 1124

Tversky A and Kahneman D, 'Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment.' (1983) 90 Psychological Review 293

——, 'Advances in Prospect Theory: Cumulative Representation of Uncertainty' (1992) 5 Journal of Risk and Uncertainty 297

Wilson TD and Gilbert DT, 'Affective Forecasting' (2003)

Wyden R and Dorgan B, 'Wyden, Dorgan Call For Immediate Halt to Tax-Funded Terror Market' (*U.S. Senate Press Release*, 2003) <http://wyden.senate.gov/media/2003/print/print_07282003_terrormarket.html>

Yi SKM and others, 'The Wisdom of the Crowd in Combinatorial Problems' (2012) 36 Cognitive Science 452

Treaty on the Functioning of the European Union, Part Three, Title XX, Article 191

APPENDIX: THE BIAS LIST FOR MINDFUL FORECASTING (BLMF)

Name of Bias	Expression of Bias	Forecast-related example
Self-confirmation bias	Tendency to search, focus and remember information that confirms one's preconceptions.	Searching for information that confirms one's forecast, before informing oneself on the topic.
Availability bias	Tendency to overestimate the probability of events with a greater "availability" in memory.	Basing a forecast by readily available information in biased or populist media outlets.
Representativeness bias	Tendency to judge a likelihood of an outcome based on its similarity to an exemplar.	Basing forecasts on a "stereotypical occurrence".
Ambiguity effect	Tendency to avoid options because of missing information.	Not making a certain forecast due to a lack of education or information available on the topic.
Attentional bias	Tendency for one's perception to be affected by one's recurring thoughts.	A forecaster's mood influencing what he attends to and perceives while reviewing information.
Backfire effect	Tendency to disconfirm evidence by strengthening one's previous beliefs.	Forecaster's being motivated to find evidence for a certain forecast.
Base-rate fallacy	Tendency to ignore generic, general information and focus on information that pertains to a certain case.	Forecaster's making bold predictions that ignore the base-rate occurrence of certain events.
Declinism	Belief that society is trending towards a decline.	Forecasters predicting overly-negative outcomes.
Gambler's fallacy	Tendency to think that future probabilities are altered by past events, when the events are in fact independent.	Forecasters predicting future events, based on past, non-related, events.
Anchoring	Heavily basing decisions on the first piece of information acquired on a subject.	Forecasters heavily basing their prediction on the first information they acquired on a topic.
Bandwagon effect	Tendency to believe in certain things because most people have the same belief.	Basing forecasts on folk knowledge.
Belief bias	Tendency to evaluate an argument based on the believability of the conclusion.	An issue for people that are evaluating a forecaster's prediction.
Bias blind spot	Tendency to see oneself as less biased.	Forecaster not being aware of their biases.
Congruence bias	Tendency to test for one hypothesis, while not testing other hypotheses.	Forecasters not examining how alternative predictions might be true.
Conjunction fallacy	Tendency to assume the a specific condition is more probable than a general condition.	Making forecasts that don't take into account the breadth of an outcome.

Curse of knowledge	Tendency of better informed people to find it difficult to think from a perspective of a lesser-informed person.	Forecasters not taking into account that the people they are forecasting about are not all knowing.
Distinction bias	Tendency to view two options as more dissimilar when evaluating the simultaneously.	Forecasters seeing an alternative prediction as more different from their chosen prediction.
Empathy gap	Underestimating the influence of feelings on people.	Forecasters not taking into account the emotional states of people while forecasting.
Focusing effect	Placing too much importance on one aspect of an event.	Basing a forecast on one piece of information.
Hot-hand fallacy	Belief that a person who has experienced success will continue on experiencing success.	Forecasting the future success of an organisation exclusively on its past success
Normalcy bias	Not planning for disasters that have not happened before.	Not predicting future catastrophic events, because they have not occurred in the past.
Optimism bias	Tendency to be over-optimistic about future events.	Not predicting a war, despite of the evidence.
Stereotyping	Expecting a member of a certain group to be more likely to perform a certain behaviour.	Making predictions based on stereotypes.
Third-person effect	Belief that mass media has a greater effect on others than on oneself.	Not being aware of the influence of forecaster's sources have on him/her.
Frame blindness	Lack of attention to the framing or definition of the problem or question, leading to unidentified options.	Forecasters discarding possible outcomes due to the phrasing of the prediction.
Non-regressiveness bias	Insufficient allowance for regression towards the mean.	Forecasters giving too much weight to extreme events explainable by statistical fluctuations.