Improving Forecasting for Foreign Policy*

Identifying drivers of forecasting accuracy in a forecasting tournament

Alexander Sacharow

April 28, 2017

Abstract

In this paper, I explore the drivers of forecasting accuracy for geopolitical events with the help of a forecasting tournament. The paper analyses the responses of the participants in order to test and evaluate several explanations of successful forecasting. More specifically, it looks at (1) measurable characteristics of forecasters, (2) the decision environment in which a forecast is made and (3) minimal interventions aiming at improved forecasting judgements. My findings are that intelligence is [good/bad] indicator of forecasting success [reinforcing /contradicting] prior findings on geopolitical forecasting. I also find that moral judgments [do (not) interfere / are unrelated] to forecasting accuracy. Regarding the decision context, this research provides evidence for a [direct correlation between / decreasing marginal return on] time used for forecasting [and / in terms of] forecasting accuracy. Similarly, participants working in teams were found to be [more/less] accurate forecasters. Finally, the forecasting competition indicated that small interventions in form of analytical guides can [not] improve forecasting judgements. The results were derived from a security policy forecasting tournament which took place from February to April 2017 and which had more than 200 participants, comprised out of university students with a strong interest in the field, paid online respondents and voluntary online users.

Contents

*Replication files are available on the author's Github account

1 Introduction

Foreign policy makers, just like other decision makers, have to constantly think about the future and how it can unfold. They need to have an idea of possible outcomes and their likelihood in order to make their decisions. For this, they rely mainly on explicit or implicit forecasting judgements, either by themselves, their advisors or some outsiders. However, forecasting geopolitical trajectories in an uncertain world has proven to be a challenge. Too often forecasts are flawed and therefore unreliable for decision makers. In order to improve future-oriented foreign policy making, a better understanding of successful forecasting is crucial.

There are several ways how this can happen. First, a better understanding leads to better forecasts which are a prerequisite for pro-active policy. Foreign policy is often described as dominated by reactive policy making. In order to break this pattern, more reliable methods for forecasting possible futures are essential. Second, it will draw the attention to ill-conceived assumptions underlying today's decision-making and thereby offers a chance to address them appropriately. Third, a better understanding of forecasting is necessary to transform institutions tasked with forecasting. Knowledge about individual differences between forecasters can be used to staff and structure such agencies. A better understanding of the decision environment will allow changing this environment to make it more suitable for forecasting. Likewise, tested decision aids can be used to counter common decision-making flaws made by forecasters. This research focuses primarily on the third point, but indirectly it does also contribute to the other points.

In this paper, a forecasting tournament is used to identify drivers behind accurate forecasting. The tournament is used both to test prior findings in the literature and to deepen the knowledge on forecasting accuracy by testing new hypotheses.

Compared to other forecasting modes, a tournament has the advantage to force forecasters to make precise predictions and it evaluates these predictions based on their track record. This reduces the level of ambiguity in forecasts, which is common to many other forms of forecasts, and measures the quality of forecasts against what they actually attempt to do.

However, it does also highlight the limitations of this research: A forecasting tournament defines successful forecasting in terms of accuracy. Participants have to specify the likelihood of a particular event in a given time frame in probability and they perform well if they choose probabilities close to the truth. Other indicators for successful forecasting are sidelined, e.g. identifying relevant possible future events or specifying the impact of certain future events.

This paper is structured as follows: First, the literature on forecasting and in particular forecasting competitions is reviewed and critically discussed. In the second section the hypotheses of this research are presented and their theoretical as well as empirical background is discussed. Third, the set-up of the research design is explained and some crucial design choices are reviewed. Fourth, the results of the research are presented. Finally, the research in general is discussed and some

policy recommendations are drawn.

2 Literature Critique

In order to forecast one has to make a basic assumption: Forecasting is possible at all. Not everyone, however, agrees to this. Forecasting sceptics emphasize the fundamental uncertainty of the future and assume successful predictions about world politics are ultimately grounded in luck (Almond, Genco 1977; Taleb 2007). They don't claim that forecasting in general is impossible, but in their view the fundamental problems about foreseeing the future are particular strong in the field of international politics. And they have a point, as foreign policy has many unfavorable conditions which have proven to be unfavorable for forecasting: The environment is dynamic, most events are essentially unique, feedback on forecasts has long delays, there is a lack of empirical tested decision aids and a strong reliance on subjective judgments (Shanteau 1992). There are even indications that the quality of judgements gets worse with professionalization as intelligence experts specialized in forecasting seem to exhibit even more decision-making biases than college students (Reyna et al. 2014).[1]

But others scholars are more optimistic about the prospects of forecasting in the field. In their view, forecasting foreign policy is still in its infancy which is partially attributed to the dominant role explanation enjoyed in the past among international relation scholars (Ward 2016). This is, however, gradually changing as a result of more forecasting-oriented research. On the individual level, the research has shown grave differences between individual forecasters, making forecasting not only a question of how to forecast but also of who is forecasting (Bueno de Mesquita 2009; Mellers et al. 2015; Tetlock 2005) and that forecasting can be further improved by appropriated training (Mellers et al. 2014). There has also been a rise in the number of methods available and the data used to generate more sophisticated forecasts for world politics (Dhami et al. 2015; Ward 2016).

This research is based on the presumption that forecasting is possible, but to a limited extend. Forecasting will never produce a fully certain prediction of the future and some events and aspects will remain beyond the forecastable. However, there are aspects which can be improved and the level of uncertainty can be reduced in a systematic and reliable manner.

The question is how to improve foreign policy forecasting in a meaningful way and what method to use for it. Generally, a wide variety of approaches is available:

Starting from the simplest and most common one: Intuitive predictions. These are statements people make out of their head without recourse to a systematic methodology. The good thing about these predictions is that they can be applied to most topics at almost no cost. Unfortunately, they have a record of being inaccurate. Take for example affective predictions (Wilson, Gilbert 2005)

[1] Which has been attributed to bad habits developed in their working environment.

which rarely match the actual experience (Schkade, Kahneman 1998) or probability judgments which have been shown to be susceptible to biases (Kahneman, Tversky 1974). The most simple statistic models have shown to outperform intuitive predictions in various domains like university admission or parole violations (Dawes et al. 1989; Swets et al. 2000). More recent research has demonstrated that expert prediction exhibit the same problems. In the political sphere, by tracking expert statements for more than 20 years Tetlock (2005) has argued expert predictions are often as accurate as a "dart-throwing chimpanzee". Similar results can be found with experts in other fields, e.g. climate science (Green, Armstrong 2007). One of the reasons for this discrepancy is that forecasting and analysis require different set of skills and there is no reason that experts combine both of them. The other reason is that experts tend not to apply the same rigidity to forecasting statements when asked about it than to their, often written, work. This does not mean expert opinions should be disregarded, but for forecasting purposes they should be treated with care.

One approach to solve the problems of inaccurate forecasting is to rely more on quantitative methods, as it is common in other disciplines like meteorology or economics. Statistical forecasting has been discussed for long in the sphere of international relations (Choucri 1974) and the field is clearly on the rise (Ward 2016). Well known applications are election forecasting models (e.g. Lewis-Beck 2005; Norpoth, Gschwend 2010) or the work of the Political Instability Task Force (Goldstone et al. 2010), which led to the Integrated Crisis Early Warning System (ICEWS) Project. The forecasts are generally based on measurable input variables ranging from economic, media to political indicators. Using past data the models are calibrated and then used for extrapolating into the future.

But there are also severe limitations to this approach. First, statistical models are largely limited to quantifiable events which can be grouped by their similarity. Hence, possible forecasts are, for example, about the outbreak and scale of violence or protests. Many events in foreign policy, however, are at least to some degree unique. Take for instance a court ruling of the International Court of Justice on a specific matter. It is hard to impossible to build a statistical model for such types of events. Hence, there are often no statistical models available to forecast relevant events or important information has to be neglected in order to make events predictable with quantitative models. Moreover, in many cases the necessary input data for quantitative models is scarce, not available or too expensive to gather. This limits the usefulness of quantitative models for policy makers. However, if there are statistical forecasting models with a good track record available, using them is surely a promising approach. The focus of this research is, however, more on geopolitical events for which quantitative models so far haven't been able to produce useful forecasts.

Prediction markets are another approach for generating forecasts which has been discussed on the literature (see Wolfers, Zitzewitz 2004). Economists see them as an efficient aggregator of information by encouraging market participants to use various information sources and by exploiting the wisdom of the crowd effect. This effect was famously described by Galton (1907), who observed that average of all estimates of the ox weight at an exhibition was much more precise than any

individual estimate. Markets essentially attempt to incorporate this effect. A well-known prediction market example is the Iowa Electronic Market , which has been used mostly to forecast elections in the U.S. But markets have been also used to forecast other political events as well, e.g. the outcome of referendums or even terrorism.[2]

For several reasons forecasting markets never really took off and still are of limited utility for policy making. One the hand this is due to the operating of prediction markets. One problem with market predictions is that they have shown to exhibit a favorite long shot bias. Small probabilities are overvalued and near certainty undervalued. This has been extensively discussed in the case of horse races, where people over proportionally bet on underdog horses (Thaler, Ziemba 1988). It was also shown to be a problem in financial markets (Bates 1991; Rubenstein 1994). Other operational problems of prediction markets are trading by desires (e.g. Forsythe et al. 1999) and speculative bubbles. However, overall these problems tend to decrease with market size and sophistication.

More problematic for prediction market are its practical limitations: First, markets are not always feasible or even desirable. Take for instance the case of terrorism, where betting on these events create perverse incentives for conducting attacks. Or the case of asymmetric information, where strong insider knowledge exists and outsiders are basically discouraged to participate in the market. This is for example the case for many government decisions, where relevant forecasting information is only available to a small cycle of individuals. Second, prediction markets do often lack the necessary liquidity and number of market participants in order to produce meaningful predictions.[3]

A relatively new approach in forecasting international politics are forecasting competitions. In such a tournament participating individuals or teams are asked about the likelihood of future events. After the end of the forecasting horizon these forecasts are compared and evaluated. Forecasting competitions are some similarity to citizen forecasts, where for example voters are asked whom they consider most likely to win their constituency (Murr 2011). Forecasting competitions combine the advantages of prior discussed approaches. Like intuitive predictions, they are hardly limited in their scope.[4] They can incorporate reliable statistical models, if they are available. And like prediction markets, forecasting competition operate on the idea of aggregating the wisdom of the crowd and aggregating different information sources in order to generate forecasts.

In the field of international politics this approach became prominent by the IARPA tournament, a geopolitical forecasting competition started in 2011 by the U.S. intelligence community. Different groups of academics were invited to participate in the project and compete against each other in providing forecasts about geopolitical events. In consecutive years, the winning team demonstrated that more accurate forecasts can be achieved by using two strategies: First, the skillful aggregation

[2] There was a DARPA research project on Policy Analysis Market (PAM), it was however stopped after public criticism.

[3] Betting markets might be a bit of an exception, but so far betting has not been used much as a political forecasting tool for events other than elections and referendums.

[4] The only serious restriction is, that the events under consideration are measureable in order to make subject to a forecasting completion.

of individual judgements (Atanasov et al. 2016). Second, the exploitation of individual differences in forecasting skills and judgements (Mellers et al. 2015). The first approach will not be discussed here. Prior research on the second strategy has primarily looked at different predictors of forecasting success (Mellers et al. 2015; Poore et al. 2014). This paper is adding to this research by testing the replicability of prior findings and exploring new hypotheses.

Before I turn to the discussion of hypotheses, I want to touch upon the limitations of forecasting competitions. Limiting the understanding of successful forecasting to accuracy is their main weakness. One the one hand, forecasting is about more than just accurately describing the likelihood of events. It does also require exploring the unknown and identifying possible future events. In the forecasting competitions discussed here, this is not part of the tournament itself.[5] It ultimately requires recourse to other methods, e.g. expert opinion or strategic foresight (Popper 2009; Kosow et al. 2008; Bergheim 2009). On the other hand, accuracy might not be the most relevant quality of forecasts for policy decision making. For this other dimension like the impact or the possibility of early action might be more important.

There are, however, several reasons why accuracy should be considered of key importance for forecasting. First, the likelihood of future event is essential for decision making from a normative point of view. Decision theory under uncertainty, and in particular its most used approaches (expected utility theory (Neumann, Morgenstern 1944) and subjective expected utility (Savage 1972)) presuppose the decision maker has an idea of the likelihood of the outcomes. Moreover, in practice decision makers tend to ask for a likelihood assessment when presented with possible future scenarios.[6] Second, improving accuracy is a process which has positive effects on the other aspects of forecasting (Tetlock, Gardner 2015). It creates good incentives for better forecasts in a broader sense by reducing ambiguity in forecasts, encouraging learning from mistakes and forcing the organizers of a forecasting competition to explore the "unknown".

3 Theory and Hypothesis

Forecasting competitions can be used to improve forecasts by identifying good forecasters, the environment in which they make good decisions and decision aids which can improve the quality of the judgement. Following this idea, the factors discussed in the paper can be divided along three categories: dispositional, environment-related and intervention. Dispositional factors are characteristics of the forecaster like abilities. They don't have to be fixed, but they should be at least stable on the short-term and measurable by a third party. Environment-related factors refer to the decision situation, e.g. the available time or whether the judgement was made alone or in a team. Intervention refers to treatments by outsiders, which aim at improving the quality of a decision.

[5] Here this is done by the tournament facilitator, which is discussed in more detail in the section on Security Policy Forecasting Tournament [CROSS REF]

[6] Observation made by scenario and foresight experts which they expressed to me in a private conversation.

Accepting the idea of individual differences in forecasting skills implies that there are individual dispositions which have direct implications on forecasting. In order to identify good forecasters it is therefore essential to know what characteristics can predict successful forecasting. *The knowledge about these dispositional factors can be used to staff institutions tasked with forecasting or distribute tasks between different individuals or agencies.*

A good starting point for this is intelligence. It has been shown to be a good predictor for many other things like job performance (Ree, Earles 1992; Schmidt, Hunter 2004), socio-economic status (Strenze 2007), academic achievement (Furnham, Monsen 2009) and decision competence (Del Missier et al. 2012; Parker, Fischhoff 2005). Therefore, it can also be assumed to be a valuable predictor for forecasting success.

Intelligence can be defined in different ways. Generally, it is conceptualized either one-dimensional or multidimensional. In the one-dimensional approach a meaningful intelligence measure can be collapsed into one variable, while in the multi-dimensional approach intelligence is understood as a collection of different abilities.

Mellers et al. (2015) have identified three relevant aspects of intelligence for geopolitical forecasting. They are (1) inductive reasoning (e.g. linking current problem and historical analogy), (2) cognitive control (e.g. override seemingly obvious but incorrect responses and engage in more prolonged and deeper thought) and (3) numeric reasoning (understanding mathematical dimension of a problem). In their research the different aspects were correlated to more accurate forecasts and they are correlated between each other. The first finding shows that intelligence measures can be used for predicting forecasting success and the second point supports the view that a one-dimensional intelligence concept is sufficient in this context. Similar results were derived by Poore et al. (2014), who found a strong correlation between various measures for analytical abilities and forecasting accuracy. Therefore it is reasonable assume these findings will be confirmed by the security policy forecasting tournament:

Hypothesis 1a: More intelligent individuals are more accurate forecasters

To test the hypothesis, it needs to be clarified how intelligence is measured. Various psychometric measures are available. Mellers et al. (2015) for example use three different methods: The Ravens Advanced Regressive Matrices method (Bors, Stokes 1998), the Cognitive Reflection Test (CRT) by Frederick (2005) and a combined numeracy scale from Lipkus et al. (2001) and Peters et al. (2006). Poore et al. (2014) used self-reportet SAT scores, sample GRE/SAT questions, subjective numeracy (Fagerlin et al. 2007) and – like Mellers et al. – the CRT.

This paper uses a test which was not used for forecasting accuracy yet: The Berlin Numeracy Test (BNT) by Cokely et al. (2012). It is a relatively new psychometric scale that was in particular developed to access statistical numeracy and risk literacy. This makes the test in particular suitable for probability judgements. It is also likely to perform better than numeracy scale used by Mellers et al. (2015) which could not appropriately differentiate between the respondents because it especially

suited for university students. Another reason for the test its length: It consists only of four questions, making it suitable for a one off questionnaire.

Intelligence is a common measure, but by far not the only dispositional factor of interest. Other factors considered in this context include personality and cognitive styles (Mellers et al. 2015; Poore et al. 2014). It has been found that factors like openness to new experiences and active open-mindedness are good predictors for forecasting success. To keep the forecasting tournament within reasonable time limits for the participants, they are not tested for their replicability here. Instead, the other factor scrutinized is one which has not been intensively tested before. It was raised by Tetlock, Gardner (2015, p. 226) in a side note: The interference of moral judgements with analytical judgments. Tetlock claims in his book that superforecasters can separate their analytical judgement from their moral judgements. This plays, for example, a role when forecasts are made about events on which the forecasters have a strong moral position. In this case many people mingle the desirability of an outcome with the likelihood assessment. For example, if a person holds a strong moral opinion on the Syrian government and has a strong desire for it to be replaced by a more human-rights oriented government, Tetlock presupposes this person to skew his or her probability judgment on the fall of the government towards the desired outcome.

This idea is similar to the socially desirability bias where individuals over-report characteristics about themselves which they consider socially desirable (Dalton, Ortegren 2011) or to trading according to desires in prediction markets, where personal political preferences affect buying decisions. Analogously, people are expected by Tetlock to overrate the likelihood of events they deem desirable.

The question is, whether this is actually true. The straight forward way to test this would be to ask respondents about the desirability of events and see whether they are correlated to their probability judgements. But this approach does not help us further here.

Ideally, we would like to predict the forecasting success of individuals without knowledge about their personal opinions. This is, on the one hand, for practical reasons, as it is both unfeasible to ask each forecaster every time about their personal view and, on the other hand, undesirable as forecasters should not be forced to reveal their personal opinions in a politicized environment. Moreover, if shown to be correct, this would imply ideal forecasters not to have opinions. This is neither desirable nor likely to be correct. Instead, as Tetlock already suggested, it is more plausible to ask whether individuals can separate their moral judgement from their analytical assessment.

In order to investigate this, we need a measure which can be assessed without reference to the actual event and which has proven to be a reliable measure in prior research. Since Tetlock assumes the underlying issue is the morality of the event, moral competency could be an appropriate measure.

Moral competency "the ability of a subject to accept or reject arguments on a particular moral issue consistently in regard to their moral quality even though they oppose the subject's stance on that issue" (Lind 2008, p. 200). It can be contrasted with opinionated judgements, which are intuitive

and emotional reactions to the content at hand and much of what Tetlock understands by moral assessment. Moral competency hence captures the idea that individuals are willing to consider counterviews despite their own, potentially strong, view on the issue.
In this regard moral competency has similarities to the active open-mindedness measure (Baron 2007), where respondents are asked on whether they would consider other opinions. But it goes a step further: Instead of relying on a self-assessment, it actually tests whether different arguments are considered in the face of a moral issue. The hypothesis is therefore:

Hypotheses 1b: More moral competent individuals are more accurate forecasters

Moral competency can be measured with the moral competency test (MCT). The test confronts respondents with two moral dilemma situations. To each one they then have to answer 12 questions on different justification for the described acts. The answers are used to compute a competency score for each respondent. The score is not based on correct and wrong answers, but reflects a ratio between different parts of the answers. [7]

Individual differences in forecasting accuracy are also determined by differences in the decision environment. Unlike early decision theory assumed, empirical research has shown that context matters (e.g. Kahneman, Tversky 1974). The list of possible factors is long, ranging from the number of forecasters, the incentives to the opportunity of deliberative practice (Ericsson et al. 1993; Kahneman, Klein 2009; Arkes 2001). In this paper only the simplest environmental factor will be tested: Time used for answering the forecasting questions.

Time used for forecasting is ultimately a choice of the forecaster. It can, however, be influenced by explicitly making time slots available or freeing forecasters of other tasks. To justify such choices, it would be necessary to know the relation between time and forecasting accuracy. There are good reasons to belief that more time will also lead to more accurate forecasts. First of all, time is necessary in order to go beyond intuitive thinking and to engage in analytical thinking about the question at hand. The two different ways of thinking are often described as 'system 1' and 'system 2', where system 1 stands for fast intuitive judgements and system 2 for slow reflective thinking (Evans, Stanovich 2013; Kahneman 2013). There is also a second reason why more time might lead to more accurate forecasts: Spending time on a question does allow gathering more information. Hence, the use of time should indicate whether a decision was informed or not.

However, there are two reasons why this might not be the case. First, the forecasters might use the time for other things. Haran et al. (2013) for example have argued that acquiring new information

[7] Based on the instructions of the test (Lind 2008) I reengineered the underlying formula for the score:

$$c = \frac{\frac{1}{4}\sum_{i=1}^{6}(\sum_{j=1}^{4} x_{ij})^2 - (\frac{1}{24}\sum_{i=6}^{6}\sum_{j=1}^{4} x_{ij})^2}{\sum_{i=1}^{6}\sum_{j=1}^{4} x_{ij}^2 - (\frac{1}{24}\sum_{i=6}^{6}\sum_{j=1}^{4} x_{ij})^2} \cdot 100 \quad (1)$$

where $i \in 1,...,6$ stands for the stage and $j \in 1,...,4$. for the section (pro-Worker, contra Worker, pro-Doctor, contra-Doctor). The score is between 0 and 100 and is sometimes categorized as follows: very low (1-9), low (10-19), medium (20-29), high (30-39), very high (40-49) and extraordinary high (above 50) (Lind 2008, p. 200). For further information check the associated website.

depends on other characteristics like active open mindedness. But this is unlikely to be a problem as participants in this case would have no reason to spend more time on the questions and move on in the forecasting tournament survey. Second, the forecasters might have different levels of pre-knowledge. Hence, some participants might need more time to grasp the context of questions while others can rely on their extensive political pre-knowledge. As will be later described in more detail, this was counteracted by ensuring a wide span of questions. Moreover, going back to the first explanation for the link between time and accuracy: Even individuals with pre-knowledge will have to switch between the two mental modes, which should again be captured by the time spend on the questions.

But the relationship between time and accuracy is unlikely linear. From the view of a mental system shift, there is no theoretical reason why more time should increase accuracy once the shift of mental systems took place. From the informational point of view, over time the value of new information decreases as it will have less and less implications for the judgement and the costs of gathering will increase as it will be harder to find new additional information. For this reason, it is reasonable to assume that the marginal value of more time decreases:

Hypothesis 2: The marginal added value of spending more time on forecasting decreases as more time is used for forecasting

In order to verify the hypothesis, the time participants used to answer the forecasting questions is measured. The validity of the measurement is ensured by treating other questions asked over the course of the forecasting completion in different sections and thereby excluded from the time measurement for this hypothesis. Moreover, the time used for forecasting is cross-checked with self-reported time use.[ˆ[note] Reasoning: The one-off survey format does allow better to monitor the time used for actual forecasting than the long-term forecasting tournaments where forecasting gets mixed with other activities during the day Hypothesis will also indicate how serious the competition was taken by participants and whether the data hence can be used to draw conclusions for professional forecasters]

Finally, differences between individual forecasters can be the result of outside interventions. In forecasting this could be achieved by treatments improving analytical judgements (Soll et al. 2015; Larrick 2004). Possible interventions include minor decision aids (Kretz 2015), feedback (Benson, Önkal 1992), exposure to multiple perspectives (Ariely et al. 2000; Herzog, Hertwig 2009), exposure to historical analogies (Lovallo et al. 2012), decomposition of problems into subsets (Fischhoff et al. 1978), explicit consideration of contradictory evidence (Koriat et al. 1980) and probabilistic training (Mellers et al. 2014). They can be tested in a forecasting tournament. Mellers et al. (2014) for example tested the effect of scenario and probabilistic trainings and found them to have a long-term positive effect on forecasting accuracy.

However, for the forecasting tournament in this paper the focus will be rather on minimal interventions as the scale of the tournament is rather small. It is therefore more suited to test mild

interventions aiming at debiasing. This could for example be achieved by providing forecasters with analytical tools. Kretz (2015) has done some research on mild decision aid interventions and found that most analysts tend to disregard decision aids, at least after some time. Kretz sees the reason for this in the additional mental capacities needed to apply the decision aids, which distracts the analysts from the actual problem at hand.[8] In his research only the mildest intervention proved to have a significant effect on the judgement quality. Following this idea, the paper tests a minimal intervention in the context of forecasting and tests the following:

Hypothesis 3: Mild interventions encouraging the use of an analytical approach increase the accuracy of forecasting

In order to test this, all participants are assigned randomly, in about equal shares, to a treatment or control group. The treatment group is provided with a short (ca. 150 words) information about the way many successful forecasters would approach the forecasting question. The guide is encouraging the forecasters to think in terms of inside and outside view (Lovallo et al. 2012), to find appropriate reference classes for the question and illustrated with an example.[9] At the end of the guide, they were asked to fill out a small check box on whether they have read the guide.

4 Security Policy Forecasting Tournament

The forecasting tournament took place from 06.-12. February 2017 and forecasters were asked to consider possible events happening between 12th of February the 24th of April 2017. The forecasters came either from the Master of International Relations program at Hertie School of Governance Berlin or were recruited externally via mailing lists of relevant study programs, associations working the field of international relations or by word of mouth. The students had to do the tournament as a homework while the others participated voluntary. The group was further complemented by forecasters from Amazon Mechanical Turk, who were paid for their participation. In total, 214 forecasters provided valid answers, they had an average age of 31 and 45,3 percent of them were female.

```
##
## % Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas
## % Date and time: Di, Mrz 28, 2017 - 20:36:10
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lccccc}
## \\[-1.8ex]\hline
```

[8] [NOTE]Limits of de-biasing: http://sourcesandmethods.blogspot.de/2014/03/reduce-bias-in-analysis-why-should-we.html

[9] The full text of the guide is available in the online appendix

```
## \hline \\[-1.8ex]
## Statistic & \multicolumn{1}{c}{N} & \multicolumn{1}{c}{Mean} & \multicolumn{1}{c}{St. Dev.} &
## \hline \\[-1.8ex]
## bnt.s & 214 & 2.056 & 1.393 & 0 & 4 \\
## mct.c & 185 & 0.163 & 0.099 & 0.000 & 0.558 \\
## time.fq.sec & 212 & 12.221 & 19.965 & 0.708 & 184.908 \\
## Duration.min & 214 & 201.031 & 707.919 & 3.167 & 4,640.800 \\
## age & 214 & 31.439 & 9.170 & 19 & 68 \\
## \hline \\[-1.8ex]
## \end{tabular}
## \end{table}
```

Statistic N Mean St. Dev. Min Max

bnt.s 214 2.056 1.393 0 4 mct.c 185 0.163 0.099 0.000 0.558 time.fq.sec 212 12.221 19.965 0.708 184.908 age 214 31.439 9.170 19 68

The forecasting tournament was conducted with a one off online survey. The survey consisted of three parts: First, the forecasters were asked question batteries of psychometric measures on intelligence and moral judgement. In the second part, the participants answered 24 forecasting questions on various security policy related events. In each forecasting question the participants were asked to provide a judgement of how likely they thought the event is, expressed in probability. Finally, in the third part, forecasters reflected upon their forecasting and provided some demographic information about them.

The forecasters were informed that their background information will be handled confidentially and in case of the university group that their forecasting judgments will be accessible to their fellow students. The second had the intention to incentive the students to give serious consideration to their answers by creating a competitive environment. In case of the voluntary participants this was less important as their participation already indicated intrinsic motivation. For the Mechanical Turk users the survey had attention checks, which they had to pass in order to receive the payout. All participants were also explicitly informed to use all information they want and spend as much time on the questions as they need. The participants were not informed about whether to work individually or together as this lies outside of the control of the research design, but they are asked in the final section whether they actually did so.

The questions were all related to security policy and selected in a multi-stage procedure. In the first step, I selected conflict regions which might be subject to changes in the time horizon under considerations. Then I developed draft questions based on surveying reports from international organizations, think tanks, governments, NGOs and media outlets on recent developments in these conflicts. Among these sources were reports by the International Crisis Group, the German Institute for International and Security Affairs (SWP Berlin), German Institute of Global and Area Studies

(GIGA), Brookings Institute and the Carnegie Center. Finally, the draft questions were sent to a few researchers and forecasting experts for feedback and their recommendations were integrated in the final forecasting tournament.

The questions were all binary and the possible answers "yes" or "no". Most questions covered possible events in the whole time period between February 12th and April 24th, 2017. For each question the forecasters had to specify a probability to indicate how likely they expected the event to be. One question was, for example, "Will IS claim responsibility for another attack with a truck inside the European Union by 24. April 2017?". The questions have to be precise and measureable. How difficult this is, one can see by the mentioned question. On March 22th, 2017 the Westminster Attack happened in the UK. The attacker used a SUV to attack and kill several people in London. The question was intended to capture such events, but the term 'truck' literally understood does not include SUVs. This is a fundamental problem about outlining events which did not happen yet: There will be aspects which were not anticipated correctly. In this case, the type of the car. How to respond in such cases? Here the event was nevertheless seen as a 'yes' reply to the question. First, the question was intended to capture such events and the use of an SUV instead of a truck does not make it fundamentally different. Second, suppose one would ask the participants whether their expectation explicitly excluded the case of SUV being used the likely answer would be no. There is no fixed rule for these borderline cases and they need to be decided on case to case basis.

Selecting the questions illustrates the distinction between two challenges in forecasting: Sampling and accuracy. In this research design sampling is done by the organizer of the forecasting competition while participants are solely dealing with the issue of accuracy. Ideally one would also include sampling into a competition and testable format, but samples of different possible future events are hard to compare and therefore they cannot easily be made part of a competition.

As a good sample of possible geopolitical events is crucial for the forecasting tournament, the forecasting questions had to satisfy a number of criteria. First, the questions should neither concern events which have almost no chance of happening nor almost certain events. It is difficult to select a highly unlikely event as they are numerous and can have all kinds of realizations. An example for such an event is the start of the Arab Uprising in 2011 after Mohamed Bouzid set himself on fire. Neither should almost certain events be subject to a forecasting competition. An example for such a question would be whether the German federal elections will take place in September 2017. To some degree such a question is a just the flipside of the highly unlikely event, but without specifying what exactly in interrupting the expectation. But again, choosing a relevant almost certain event for a forecasting competition will become an arbitrary choice. Moreover, remote and almost certain events will cause clustering of forecasts along the extreme values by participants in the forecasting competition. This would make it harder to distinguish successful forecasters from unsuccessful ones.

Second, the questions should cover various regions. The results might be biased if individual participants have special knowledge about a region which is overly represented in the competition.

To reduce the possible complications further, similar question can be avoided. In the case of the security policy forecasting tournament the question topics were constrained by had to collaboration with the security policy course and chosen along the main topics of the course.

Third, events for the forecasting competition should be relevant for policy makers and a large group of people. Relevance implies that the event has an impact on policy makers. The impact dimension excludes forecasting questions like the music played at the inauguration of a head of state. Relevance in this research was ensured by selecting events or indicators which would be discussed or used by international organizations, governments and policy-oriented research institutions.

Even though these criteria guided the selection of the questions, a few limitations had to be taken into account. First of all, language restricted the number possible events as only events for which English language reporting was available could be selected for the competition. Second, the events were chosen on the basis of possibly getting international media attention. On the one side, this reduces the barrier for participants as it limited the scope of the questions to topics they might at least generally familiar with. On the other side, it keeps the workload for tracking questions reasonable. However, this does also exclude many possible questions. For example, funding decisions in international organizations are of policy relevance and might have severe implications, but information on them are hardly available. Third, for the events should be reliable information available. In the field of security policy this can be difficult as information are inherently subject to the conflict dynamics and in many conflict areas almost any reliable information is hard to get by. Take for instance the conflict in the Democratic Republic of Congo or even Syria, where smaller incidences are rarely reported, and even if, cannot independently be confirmed.

In order to assess the quality of forecasts, they have to be scored. A common method is based on the Brier score, which was originally proposed in the context of weather forecasting. (Brier 1950) Generally speaking, the Brier score indicates the distance of the forecast to the truth. More precisely, the Brier score is the squared error of a probabilistic forecast. To calculate it, the forecast are expressed on the range between 0 (0%) and 1 (100%). The realized events are coded either 0 (if the event did not happen) or 1 (if the event did happen). For each answer option, the difference between the forecast and the correct answer is squared and added. It can be expressed with:

N stands for the number of events, R is the number of possible classes the event can fall, p is the probability forecast and o the realized outcome. The Brier score can evaluate questions with more than two possible outcomes (R>2), but in this paper only binary events are considered (R=0). The best (lowest) possible Brier score is 0, and the worst (highest) possible Brier score is 2. The Brier score is a proper scoring function which means that participant cannot improve their score by reporting a different probability from their actual belief.

In total 231 individuals participated in the security policy forecasting tournament. However, only 214 responses as for some participants there were signs of not taking the survey serious (failing

attention checks, unrealistically short time used for participation) and others only submitted their forecast after February 12th, 2017.

The security policy forecast tournament was implemented with the survey tool Qualtrics. Descriptive statistics

5 Data Overview

how representative our data already is

6 Results

What the result is of making it representative

7 Conclusion

Summary of the core finding

Further implications

8 References