METHOD ARTICLE

Improving national strategic foresight with the use of forecasting tournaments and its implications for the study of international relations

Jan Kleňha

Department of North American Studies, Charles University, Prague, Czech Republic

Abstract

Improving national strategic foresight can help the formation of more robust and informed policies, including foreign policy. Predicated upon the theory behind peer-prediction elicitation methods such as Reciprocal Scoring, we combined two foresight methods - Forecasting tournaments and a Delphi method - into a design in which a forecasting tournament predicted the results of a Delphi. Experts in a Delphi could take into account the arguments of participants from a prior forecasting tournament and thus make better-informed decisions. This methodological article aims to validate the feasibility of this design. It describes how we implemented it for identifying and prioritizing global megatrends as part of a strategic foresight project for the Czech government. We found this design practically applicable, while the forecasting tournament also seems to improve the ability of participants to predict a group consensus. Similar combinations of foresight methods could be used to enhance the study of international relations.

Keywords

national strategy, foresight, forecasting tournament, Delphi method, deliberation, prediction, consensus

Corresponding author: Jan Kleňha (klenhajan@gmail.com)

Author roles: Kleňha J: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This article is a result of the author's work on a dissertation research conducted at the Institute of International Studies at the Charles University in Prague and supported by the SYLFF program (the Ryoichi Sasakawa Young Leaders Fellowship Fund) and the project SVV by the Institut mezinárodních studií FSV UK nr. 260954/2020. The data were collected in close cooperation with the non-profit organization České priority, z. ú. during its projects FUTURE-PRO (TITDUVCR946MT01) and OPTIONS (TL04000315) that were financed as research grants by the Technological Agency of the Czech Republic (TA ČR).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article:
For printed version: Kleňha Jan. "Improving national strategic foresight with the use of forecasting tournaments and its implications for the study of international relations". *Stosunki Międzynarodowe – International Relations* 57, (2021): 71–92. Printed 2022. https://doi.org/10.12688/stomiedintrelat.17424.1.
For online version: Kleňha J. **Improving national strategic foresight with the use of forecasting tournaments and its implications for the study of international relations.** Stosunki Międzynarodowe – International Relations 2022, **2**:4 https://doi.org/10.12688/stomiedintrelat.17424.1

Introduction

In the study of international relations, it is important to understand the methods used by national governments in the development of their national strategic documents. These strategic documents are often formulated with the use of foresight. Foresight studies on a national level of governance are usually carried out with the purpose of directly or indirectly guiding the future directions of the nation's domestic and international policy, which often has clear implications on the development of international relations.

The national strategic foresight is an area of anticipatory governance that often uses the methods of crowd-wisdom aggregation to deliver robust results in highly uncertain settings.[1] The "wisdom of the crowds" is a phenomenon coined by James Surowiecki[2] and expanded by Cass Sunstein[3] in the early 2000s, but intuitively known at least since the early 20th century,[4] claiming that the aggregation of judgements often outperforms the judgements of individuals.

The benefits of using crowdsourcing methods have been shown in many other contexts, such as political elections,[5] economic forecasting[6] or public policy.[7] In the case of predicting long-term trends on a national and global level, which is highly complex and difficult, many governments and institutions tend to use smaller-scale deliberative methods, where the "crowd" consists of a group of credentialed experts from diverse backgrounds, aiming to capture a wide range of sector-specific expertise.

One of the most frequently used methods for the national strategic foresight is the Delphi method, which, however, has its limitations, such as the need for high diversity of expertise, some of which might not be properly reflected by the standard academic credentials and therefore difficult to select for when inviting experts, or the time-demanding nature of a Delphi, which might be especially concerning among scholars with higher credentials or professionals as they are likely to be time-constrained.

---

[1] P. Tõnurist and A. Hanson, "Anticipatory innovation governance: Shaping the future through proactive policy making," *OECD Working Papers on Public Governance* (2020): 44, https://doi.org/10.1787/cce14d80-en.

[2] J. Surowiecki, *The Wisdom of Crowds* (New York: Anchor, 2005).

[3] C. Sunstein, *Infotopia: How Many Minds Produce Knowledge* (Oxford: Oxford University Press, 2006).

[4] F. Galton, "Vox populi," *Nature* 75, (1907): 450–51.

[5] W. Gaissmaier and J.N. Marewski, "Forecasting elections with mere recognition from small, lousy samples: A comparison of collective recognition, wisdom of crowds, and representative polls," *Judgment and Decision Making* 6, (2020): 73–88.

[6] D.V. Budescu and E. Chen, "Identifying expertise to extract the wisdom of crowds," *Management Science* 61, (2014): 267–80.

[7] M.G. Morgan, "Use (and abuse) of expert elicitation in support of decision making for public policy," *The Proceedings of the National Academy of Sciences* 111, (2014): 7176–184.

"Forecasting tournaments" is another academically rigorous method of crowd-wisdom aggregation; it has been rising in popularity over the last decade and works to solve these very limitations effectively. Its own limitations are, however, the need for a clearly definable resolution, and the need for this resolution to happen in a relatively short-term future. Therefore, forecasting tournaments cannot be used for long-term strategic foresight without being combined with other foresight methods.

One way to combine short-term forecasting with longer-term foresight for mutual benefit is to use a forecasting tournament to predict the results of a foresight study that uses a Delphi method. Predicting the results of social science studies is one of the regular fields of application of short-term forecasting tools,[8] but its use in predicting the outcomes of a foresight study based on a Delphi, with a purpose of increasing the robustness of these outcomes has not yet been described in academic literature. This methodological approach was heavily influenced by our discussions with the international forecasting community about new methods of scoring of questions without clear resolution, which also recently resulted in the study "Reciprocal Scoring: A Method for Forecasting Unanswerable Questions,"[9] funded by the US Intelligence Advanced Research Projects Activity (IARPA) and Open Philanthropy. Our approach is predicated upon the same theory of the effectiveness of peer-prediction elicitation, while it applies it in a more specific setting closer to the standard processes used in national strategic foresight.

Exploring how to use forecasting tournaments for foresight is a research direction that is in accordance with current recommendations from the scientific community. For example The Perry World House, an interdisciplinary global policy research institute at the University of Pennsylvania recommends "launching experiments focused on different types of forecasts on which there is currently little research, including conditional forecasting and longer-term forecasts".[10]

This approach is also highly relevant to the study of international relations. The Center for Security and Emerging Technology (CSET) is, for example, regularly using forecasting tournaments to inform policymakers with geopolitical predictions

---

[8] Social Science Prediction Platform - An interview with Stefano DellaVigna, UC Berkeley Social Science Matrix (September 2020), Available at https://live-ssmatrix.pantheon.berkeley.edu/research-article/social-science-prediction-platform/.

[9] E. Karger *et al.*, "Reciprocal Scoring: A Method for Forecasting Unanswerable Questions," (October 2021), pre-print version, Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3954498.

[10] M. Horowitz *et al.*, "Keeping Score: A New Approach to Geopolitical Forecasting," *Perry World House* (2021), Available at https://global.upenn.edu/sites/default/files/perry-world-house/Keeping Score Forecasting White Paper.pdf.

such as the probability of a violent U.S.-China conflict in the South China Sea[11] or the predictions of the growth or decline of the U.S.-China Trade.[12]

The aim of this methodological article is to prove the hypothesis that it is practically feasible to use a forecasting tournament as part of a national strategic foresight study, in order to increase its robustness.

In 2021, we conducted a forecasting tournament to predict results of a Delphi during the foresight project "FUTURE-PRO: Identification of global megatrends relevant for the country's future".[13] This project was delivered by the organization České priority, z.ú., that has been contracted by the Office of the Government of the Czech Republic to create a methodology for the purpose of prioritizing national funding for research and innovation.

This article describes the practical application of the aforementioned design. The goal of this article is not to present an analysis proving this design to improve the quality of the Delphi outcomes or the outcomes of the study in which this method was used. This article should serve mainly as a guidance for future researchers to improve their initial experimental design while piloting or applying similar approaches.

Structure

The main part of the article is structured into three sections. Section 1 (Methodology) contains a review of the main benefits and limitations of the selected methods, as well as the overview of alternative foresight methods and the reasons why we have not applied these alternative methods in the project.

Section 2 (Implementation) describes in detail our application of an established methodological approach (using a forecasting tournament to predict results of a study) to predict a Delphi study identifying future global Megatrends. This section primarily deals with the specific interactions of the outputs of these two methods, their timing and other practical elements.

Section 3 (Findings) features the elaboration on the main successes and failures of the implementation, including the observed benefits of using forecasting tournaments to increase the quality of expert deliberation. Possible use cases and implications of this methodological design, if they prove efficient in subsequent research, are discussed in the conclusion.

---

[11] M. Page and A. Barker, "Forecasting Conflict in the South China Sea," *CSET-Foretell* (October 2020), Available at https://www.cset-foretell.com/blog/forecasts-south-china-sea.

[12] M. Page, "Crowd Outperforms Projections from Historical Data in Early Results," *CSET-Foretell* (March 2021), Available at https://www.cset-foretell.com/blog/crowd-performance-analysis.

[13] České priority, "FUTURE-PRO: Identification of Megatrends and Global Challenges for the Czech Republic," (July 2021), Available at www.megatrendy.cz.

Methodology

Strategic foresight

As this article discusses the methods of crowd-wisdom aggregation that may be used particularly for the purpose of strategic foresight, it is important to provide an introduction to strategic foresight as a field.

Foresight can be defined as a broad set of methods for analyzing the future for the purpose of actively shaping it.[14] Strategic foresight in general is a well-established field of research, whose applications are used (to various degrees of quality and rigour) by national governments, international organizations and private companies for strategic planning and decision-making.

The strategic foresight was originally focused on military and technological development, but in the last 30 years, it has expanded to more general societal topics such as sustainable development, social policies or infrastructure. Recently, strategic foresight has been increasingly developed in the EU, particularly within the European Strategy and Policy Analysis System (ESPAS), which regularly publishes studies aimed at identifying future challenges for EU public policies.[15]

The strategic foresight used at the national level enables the shaping of specific public policy and investment measures to maximize their long-term effectiveness. In his study, Jacobs[16] describes the advantages of so-called long-termism, whereby the government prepares its strategies and measures taking into account long-term future developments. Jacobs shows that long term strategic foresight pays off significantly as it brings large long-term benefits for moderate short-term costs.

Finland[17] or the UK[18] can be noted as the examples of countries that design their public policies to anticipate long-term future developments. According to Boston,[19] preparing for future developments in society is an integral aspect of good governance. When future developments are taken into account, the strategies as well as particular measures are significantly more effective, robust and resilient.

Only a few contemporary governments are, however, adequately equipped for conducting or systematically contracting broad foresight studies. This might be partially caused by politicians often prioritizing short-term interests and by many

---

[14] Ibid.

[15] Ibid.

[16] A. Jacobs, "Policy Making for the Long Term in Advanced Democracies," *Annual Review of Political Science* 19, (2016): 433–54.

[17] Megatrends, SITRA (2020), https://www.sitra.fi/en/topics/megatrends/.

[18] The Futures Toolkit, Government Office for Science (2017), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/674209/futures-toolkit-edition-1.pdf.

[19] J. Boston *et al.*, "Foresight, insight and oversight: Enhancing long-term governance through better parliamentary scrutiny," (2019), https://www.victoria.ac.nz/__data/assets/pdf_file/0011/1753571/Foresight-insight-and-oversight.pdf.

governments having increasing reactionary tendencies, but also by the high difficulty of this task due to its large complexity and uncertainty.

Predicting medium and long term global megatrends is often seen as the first step in prioritization of areas to focus public resources on. In predicting global megatrends, it is important for the governments to obtain relatively safe and robust information of the likely future importance of various areas of emerging problems and opportunities. This is usually done by consultative methods of expert deliberation. One often used method designed to deliver robust results is the Delphi method.

Delphi method

The Delphi method was developed by Dalkey and Helmer[20] at the Rand Corporation in the 1950s. It is widely used for debate structuring among experts and either "achieving convergence of opinion concerning real-world knowledge solicited from experts within certain topic areas",[21] or identifying recurrent dissensus and conflicting views. It is a method similar to a survey, but it is based on iterations where respondents receive feedback from the previous round and can adjust their estimate based on the estimates of other experts. Another important element is anonymity, which aims to reduce "groupthink" and the prevalence of "senior" opinions.[22]

The identification of the experts who will participate in Delphi is an important element. Gordon[23] lists several ways to identify and select relevant experts. Experts can be identified based on the authorship of a publication on a given topic. However, this eliminates experts who have not published or whose publications were not noted in the literature synthesis. It is also possible to rely on recommendations from institutions, but in this case there is a risk that only experts who are known to these institutions will be identified, creating opinion "cliques". To avoid this risk, it is possible to appeal publicly for expertise through the public media (in the media, on bulletin boards, etc.). In this way, even less known experts can be recruited. It is also possible for experts to recommend others.

Experts selected to participate in Delphi usually respond to already formulated statements. These statements are usually based on a synthesis of the literature. For the selection and formulation of statements, several principles must be followed. Statements must be unambiguous, relatively short and precise. It is also

[20] N.C. Dalkey and O. Helmer, "An experimental application of the Delphi method to the use of experts," *Management Science* 9, no. 3 (1963): 458–67.

[21] C.C. Hsu and B.A. Sandford, "The Delphi Technique: Making Sense of Consensus," *Practical Assessment, Research & Evaluation* 12, no. 10 (2007), https://doi.org/10.7275/pdz9-th90.

[22] T.J. Gordon, "The Delphi method," *Futures Research Methodology - The Millenium Project* (1994), https://eumed-agpol.iamm.fr/private/priv_docum/wp5_files/5-delphi.pdf.

[23] Ibid.

recommended not to use technical or professional terms.[24] In a Delphi questionnaire, questions are often formulated to address the respondent's knowledge of the topic, an assessment of the time horizon or likelihood of a given development, an assessment of the implications or impact of the development and an assessment of factors that may hinder or facilitate the development. A range of question types can be formulated, from closed (multiple choice, rating, ranking) to open questions.

Questionnaires in Delphi can be administered on paper or on-line. Today, the most common is on-line completion, where respondents answer the questionnaire on a dedicated website. The feedback can either be displayed immediately after the respondent has answered (Real-time Delphi) or at the end of the round (usually after a few weeks). Outputs are usually presented as the distribution of responses in the form of a graph or histogram. If experts were asked to provide arguments for their answers, the results include written text that can be analyzed by the organizer or e.g. using text mining tools.

Benefits

The two main benefits of the Delphi method are the anonymity of participants and the iterative feedback.

- Anonymity - The anonymity of participating subjects can reduce the effects of dominant individuals which often is a concern when using group-based processes used to synthesize information.[25] Anonymity also helps prevent participants from making suboptimal decisions due to being influenced by the credentials, expertise or social statuses of the others.
- Iterations - the process of feedback in multiple iterations allows and encourages the selected Delphi participants to reassess their initial judgments about the information provided in previous iterations.[26] This design allows participants not only to change their mind in the light of additional information, but also not to make decisions under pressure or other circumstances.

Other benefits stem from the use of numerical responses, which could then be statistically aggregated, and from the facilitator's ability to control various parts of the Delphi process, e.g. by providing controlled feedback between rounds or by being able to use statistical analysis techniques to further reduce the risks of the participants' pressure for conformity within the group.[27]

---

[24] České priority, "FUTURE-PRO."

[25] N.C. Dalkey and D.L. Rourke, "Experimental assessment of Delphi procedures with group value judgments," *Studies in the quality of life: Delphi and decision-making*, eds. N.C. Dalkey, *et al.* (Lexington: Lexington Books, 1972): 55–83.

[26] Hsu, "The Delphi Technique: Making Sense of Consensus."

[27] Dalkey, "Experimental assessment of Delphi procedures."

Limitations

The two main limitations of Delphi are the needs for a diversity of participants and for their strong motivation.

- Need for high diversity - Selection of participants is the most important step in the entire process of Delphi, as it directly relates to the quality of the results generated.[28] Need for a wide diversity of expertise and opinions (especially if the subject of deliberation is highly complex) is crucial and difficult to obtain in practice, usually due to the financial and time limitations. Moreover, many sector experts with credentials and expertise to be nominated to Delphi might not be very good "generalists". Naturally, they may be biased towards the field they have been working in for a long time. They are also more likely to be occupied with other projects and not allocate sufficient time to provide precise arguments. Young educated participants might be less biased, think in novel ways and have more time to conduct additional research, but they often don't have the appropriate credentials yet to be invited into an expert Delphi study.
- Need for strong motivation - Motivation of participants is the key to the successful implementation of a Delphi study and investigators need to actively ensure to maintain a high response rate throughout multiple rounds.[29] Experts need to be motivated (financially or socially) to put a relatively intensive effort into reading the inputs of others and writing thoughtful comments in multiple rounds. Moreover, strong motivation is needed especially in studies with research topics that can never be clearly resolved (e.g. what should be the national priorities or what will be the global megatrends' implications) and therefore no claims are clearly falsifiable, which does not motivate participants to be maximally correct. This is especially important in the case of high-impact decision-making, where the perceived benefits from being deliberately dishonest (e.g. prioritizing issues that one has vested interests in) might outweigh the benefits from being right (e.g. being a participant in an impactful and cited study).

There have been other observed limitations such as the risk of a "pressure to conform with group ratings"[30] but they can be mitigated by decreasing the effects of the two limitations described above.[31]

---

[28] R.C. Judd, "Use of Delphi methods in higher education," *Technological forecasting and social change* 4, no. 2 (1972): 173–86.

[29] Hsu, "The Delphi Technique: Making Sense of Consensus."

[30] B.R. Witkin and J.W. Altschuld, *Planning and conducting needs assessment: A practical guide* (Thousand Oaks: Sage Publications, 1995): 188.

[31] F. Bolger and G. Wright, "Improving the Delphi process: Lessons from social psychological research," *Technological Forecasting and Social Change* 78, no. 9 (2011): 1500–513, https://doi.org/10.1016/j.techfore.2011.07.007.

Forecasting tournaments

Forecasting tournaments is a method of crowd-wisdom aggregation for the purpose of gathering informed estimation of future developments, events, trends or outcomes.[32] This method can be classified as judgmental forecasting, unlike the methods based on statistical models or machine-learning models.[33] Forecasting tournaments, as claimed, have a "potential to improve not only the quality of political decision-making but also the public awareness and participation, and hence general trust in politics".[34] Probably the most famous practical application of this method is the Good Judgement Project developed upon the theoretical findings of Phillip Tetlock and his team,[35] which provided data to multiple subsequent studies and marked the beginning of a rapid growth of the field of forecasting especially in the area of geopolitics and international relations.

Forecasting tournaments is a method to not only effectively aggregate various inputs, but also to increase the incentives of participants to put more effort into formulating their inputs by using a combination of financial and social motivations. Participants in a forecasting tournament are motivated to create and share their predictions, opinions and sources in real-time on an on-line prediction platform with others, which increases the benefits of collaboration and dissemination of ideas within a group.

Forecasting tournaments usually use scoring methods such as a Brier score[36] to motivate participants to search for the most correct probabilistic predictions and not be overconfident. It is in each participant's interest to update their own predictions during the tournament, for example if influenced by the inputs of others. Other design adjustments can be made to further improve the process and the outcomes of a forecasting tournament, such as using a Categorical scoring rule (to motivate inputs early in the tournament) or rewarding the best comments (to motivate more sharing of information).

Benefits

Forecasting tournaments offer the same two main benefits as Delphi - anonymity and iterations. The benefits of an iteration are further amplified, as the cross-insemination of views and arguments happens not in rounds, but in real time on-line. The

[32] P. Tetlock *et al.*, "Forecasting tournaments: Tools for increasing transparency and improving the quality of debate," *Current Directions in Psychological Science* 23, no. 4 (2014): 290–95.

[33] T. Januschowski *et al.*, "Criteria for Classifying Forecasting Methods (Invited Commentary on the M4 Forecasting Competition)," *International Journal of Forecasting* 36, no. 1 (2020): 167–77.

[34] J. Dana *et al.,* "Are markets more accurate than polls?" *Judgment and Decision Making* 14, no. 2 (2019): 135–47.

[35] P.E. Tetlock, *Expert political judgment* (Princeton: Princeton University Press, 2009).

[36] G.W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly weather review* 78, no. 1 (1950): 1–3.

group discussion is usually not moderated (or moderated only for inappropriate or adversarial behavior), participants can publicly react to each other's comments and change their own forecasts at any time. As a result, there is more information sharing taking place than in Delphi. Two additional benefits of forecasting tournaments are:

- Competitiveness - The framing of this deliberation process as a "tournament" increases the incentives of participants to put more effort into providing higher-quality inputs and to be right. The process factors maintaining competitiveness such as motivation, evaluation, feedback, collaboration methods etc., need to be carefully adjusted when designing the tool.[37] Financial rewards are the most commonly used motivator, but many participants claim to be motivated rather by the opportunity to test, show and improve their forecasting skills.[38] These motivations can be utilized to further reduce the costs of the method.
- Scalability - The algorithmic and automatized forecasting tournaments allows it to accommodate orders of magnitude more participants than a Delphi, which is highly beneficial for decreasing the effects of each individual input´s and increasing the diversity of the views represented in the process. As this introduces some risks (e.g. the possibility of adversarial collaboration or reputation harms caused by disruptive participants), multiple factors such as training, expertise, general knowledge of participants, etc. should be considered during the preparation.[39]

In addition, all inputs are automatically collected on a platform that can later serve as a digital repository of opinions and resources, which may be useful for capacity building and learning purposes of individuals and the institutions, as well as for increased accountability of participants for their inputs.

Limitations

- Need for clear resolutions - To function properly, forecasting tournaments need to ask questions that will be clearly resolvable in the future, to prevent disputes about the outcomes. With the potential for disputes, participants are motivated to adjust to these risks and limit their effort.
- Need for short-term questions - The financial rewards for correctly predicting long-term questions (more than 2–3 years) are less appealing due to the increasing difficulty as well as the increasing opportunity costs of conducting proper research before making a prediction.

---

[37] B. Mellers *et al.*, "Psychological strategies for winning a geopolitical forecasting tournament," *Psychological Science* 25, (2014): 1106–115.

[38] České priority, "FUTURE-PRO."

[39] Mellers, "Psychological strategies."

These two limitations notably narrow the scope of available questions, while it is precisely the long-term questions that are at the core of strategic foresight. Long-term questions also carry larger potential for high-impacts and are interesting for policymakers and the public to discuss and to make opinions about.

Combining Delphi with Forecasting tournaments

From the benefits and limitations of both methods postulated above, it is apparent that the two main limitations of a Delphi can be mitigated precisely by the two described benefits of Forecasting tournaments. The first Delphi limitation - the need for a strong motivation - can be effectively aided by introducing the competitiveness aspect of forecasting tournaments into the process. The second Delphi limitation - the need for high diversity - can be mitigated by the scalability of forecasting tournaments to accommodate a larger spectrum of views while maintaining their effective aggregation.

The two limitations of Forecasting tournaments - the need for a clear resolution and the need for short-term questions - can, in turn, be effectively solved by combining the two methods by using a forecasting tournament to predict the results of a Delphi. The Delphi results are clear (e.g. a final ranking of priorities), delivered using a given methodology and known in a short-term future. This design may increase the cost, complexity and the length of the study, but if it, in fact, helps to deliver better results, it might still be a very cost-effective design relative to the impacts of consequent strategic decisions.

Asking participants in a forecasting tournament to predict, what will be the opinion of a different group of respondents (in this case experts in a Dephi study), is an empirically credible approach based on the theory behind peer-prediction elicitation. As empirical evidence, Karger, Tetlock et al. recently found that forecasts elicited using Reciprocal Scoring method were as accurate as those elicited with Brier score & both outperformed a control group without incentives.[40]

There are a number of design choices to be made during the implementation of this design, but the most important choice is about how much information from the forecasting tournament should be fed into the Delphi. It is important for the experts in Delphi not to see the aggregate of the predictions from the previous tournament, because they could consider it high-quality information that they cannot outperform, and therefore they could give up on doing their research and formulating their own opinions.

The Delphi experts should, on the other hand, be able to see the comments and arguments of the individual participants in the tournament, as they may contain important, yet marginal or contrarian views that the experts can then reflect in the Delphi. The comments can also contain the description of likely biases of the group

[40] Karger *et al.*, "Reciprocal Scoring."

of experts, which can help the individual experts to be more aware of them and avoid them. Participants in the tournament can try to use the "self-fulfilling prophecy" phenomenon[41] for their advantage, e.g. by writing persuasive arguments for a particular response while predicting, that these arguments will in fact influence the results of the Delphi, which might have interesting effects that are currently underexplored, but this should also introduce information rather than noise to the process.

Alternative approaches

Prediction markets

Prediction markets are a popular platform for the elicitation of incentivized crowd predictions.[42] It is a method, in which participants in the market can use their own money or credit to buy and sell shares of various predictions. Prediction markets are designed specifically to forecast events such as elections,[43] which is being experimented with by a number of existing on-line prediction markets.[44] The idea of using prediction markets to predict results of scientific studies was first introduced by Robin Hanson in 1995.[45]

In theory, prediction markets should be a highly efficient method to aggregate accurate short and medium-term predictions, but in practice, it appears difficult to motivate enough participants to ensure that the markets are liquid, which is a necessary condition for them to work.[46] Especially when the resolution of the prediction is further in the future, there is an increased chance that the resolution will be disputable or the project will cease to exist, and that the capital returns will be lower than could have been elsewhere, even if one´s prediction turns out to be correct.

In addition, people are often subject to loss aversion[47] hesitating to bet their own money, even if the odds are favorable. Specifically in the case of predicting scientific results, prediction markets were recently found to be rather ineffective.[48]

---

[41] R.K. Merton, "The self-fulfilling prophecy." *The Antioch review* 8, no. 2 (1948): 193–210.

[42] A. Brown *et al*., "When are prediction market prices most informative?" *International Journal of Forecasting* 35, no. 1 (2019): 420–28.

[43] J.E. Berg, F.D. Nelson and T.A. Rietz, "Prediction market accuracy in the long run," *International Journal of Forecasting* 24, no. 2 (2008): 285–300.

[44] For example PredictIt, Augur, Gnosis or Polymarket.

[45] R. Hanson, "Could gambling save science? Encouraging an honest consensus," *Social Epistemology* 9, (1995): 3–33.

[46] R. Hanson, "Decision markets for policy advice," *Promoting the general welfare: American democracy and the political economy of government performance* (2006): 151–73.

[47] A. Tversky and D. Kahneman, "Loss aversion in riskless choice: A reference-dependent model," *The quarterly journal of economics* 106, no. 4 (1991): 1039–1061.

[48] D. Viganola *et al.,* "Using prediction markets to predict the outcomes in DARPA's Next Generation Social Science program," *The Royal Society Publishing,* http://doi.org/10.1098/rsos.181308.

These limitations suggest that in the near-term future, even moderately subsidized prediction markets can be outperformed by different approaches, which is why we have not applied this method.

Surprising popularity

Surprising popularity is an interesting academic concept of crowd-wisdom aggregation, developed by Dražen Prelec in 2007, based on the approach of asking participants to respond and also to predict the average responses of others, and then "selecting the answers that are more popular than people predict".[49] The main benefit of this novel method is that it can detect cases in which the majority of participants are wrong in their responses, which Delphi or Forecasting tournaments cannot.[50]

It also does not require future resolution of the questions, which is beneficial for the purpose of strategic foresight. On the other hand, the method consists of distributing questionnaires without information sharing between participants, and there is no rigorous evidence yet for the positive effects of using Surprising popularity in combination with other methods of deliberation. This is why we have gathered data during the FUTURE-PRO project to explore this method further, but have not applied it in this study.

Scenario planning combined with forecasting

Scenario planning is another standard foresight method that combines facts with identified driving forces to create future scenarios. This method has its own limitations such as excessive optimism about certain scenarios, an over emphasis on unlikely events, and over relying on historical precedent,[51] which might be mitigated by probabilistic forecasting. This is a very recent approach that is being developed by the scientific community behind forecasting tournaments, most notably The Cultivate Labs, the CSET and the team of Dr. Phillip Tetlock. It combines probabilistic forecasting with Scenario planning, hoping that "this holistic method would provide policymakers with both a range of conceivable futures and regular updates as to which one is likely to emerge".[52]

---

[49] D. Prelec, H.S. Seung and J. McCoy, "A solution to the single-question crowd wisdom problem," *Nature* 541, (2017): 532–35.

[50] W. Chang *et al.*, "Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments," *Judgment & Decision Making* 11, no. 5 (2016).

[51] D. Erdmann, B. Sichel and L. Yeung, "Overcoming obstacles to effective scenario planning," *McKinsey Quarterly* 55, (2015): https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/overcoming-obstacles-to-effective-scenario-planning#.

[52] J.P. Scoblic and P.E. Tetlock, "A better crystal ball: The right way to think about the future," *Foreign Affairs* 99, (2020): 10, https://www.foreignaffairs.com/articles/united-states/2020-10-13/better-crystal-ball.

In the process of "Strategic Question Decomposition", future scenarios are broken down into pivotal factors and then individual falsifiable signals, that can then be more effectively forecasted.[53] We have not used any aspects of this method as it is still relatively early in the development, but it seems to be a promising approach that could be applied in national strategic foresight as well in the future.

Implementation

In 2021, we organized a forecasting tournament to enhance the standard application of Delphi in the aforementioned project FUTURE-PRO. The core of the project was a Delphi with 24 participating experts with a pre-designed diversity of academic backgrounds, who were presented with 18 areas of global megatrends and were asked (during three rounds, among other questions) to rate (on a scale 0–3) their agreement with this statement: *"The area will have a very significant impact on the quality of life in Czechia in the next decades and, therefore, public funding should be preferentially allocated to understanding it and addressing it."* The aggregate of the final ranking of this question was used as a resolution to the forecasting tournament. The full report of the project with technical details regarding the experiment and its methodology is available at www.megatrendy.cz.[54]

The forecasting tournament was designed to take place before the Delphi and involved 238 forecasters, who had earlier passed a 1.5 hour on-line calibration training with a quiz at the end. This training was focused on explaining the basics of working with probabilities and the methods of properly estimating own confidence, which should result in making well-calibrated predictions. Participants were anonymous - each participant was instructed to choose a name of any foreign city as an identity, which was then displayed on a forecasting platform. We used a forecasting platform developed by Cultivate Labs,[55] which is being used by international forecasting tournaments such as The Good Judgement Project, projects run by the Center for Security and Emerging Technology or The British National Intelligence project "Cosmic Bazaar".[56]

The recruitment, which resulted in the 238 participants, was targeted mainly at students, PhD students and university scholars in the Czech Republic. The group was relatively young and highly educated - 59.3% of participants were under 35 and 68.2% of participants had a master's degree or higher. The group of participants

[53] A. Siegel, "Tracking the Outcome of Strategic Questions with Crowd Forecasting," *Cultivate Labs blog* (2021), https://www.cultivatelabs.com/posts/tracking-the-outcome-of-strategic-questions-with-crowd-forecasting.

[54] České priority, "FUTURE-PRO."

[55] Cultivate Labs, www.cultivatelabs.com (accessed October 2021).

[56] "How spooks are turning to superforecasting in the Cosmic Bazaar," *The Economist*, April 17, 2021, https://www.economist.com/science-and-technology/2021/04/15/how-spooks-are-turning-to-superforecasting-in-the-cosmic-bazaar.

was diverse with regards to expertise. The most common professional focus of respondents was Economics and Business (24x), Computer and Information Sciences (23x), and Political Science (16x), Physical Sciences (12x), Mathematics (8x), Legal Sciences (7x), Other Social Sciences (7x), Sociology (6x), Psychology and Cognitive Sciences (5x), Education (5x) and Biological Sciences (5x). 16 other Fields of Research and Development (FORD) disciplines were represented. 34% of participants did not answer.

All participants were trained in forecasting prior to participation and were familiarized with the technical interface for the forecasting tournament. Apart from this experiment, they used the same forecasting platform to participate in a larger, three-month long forecasting tournament OPTIONS[57] focused on short-term questions mostly related to public policy development in the Czech Republic. The tournament as well as the questionnaire and all the provided materials were in Czech language, the whole experiment lasted for two weeks and took place in April 2021 (Figure 1). A translation of the questions can be found in *Extended data*.

Questionnaire

The questionnaire was administered for research purposes. It was mandatory and was answered by 238 participants. In its introduction, participants were provided with the links to 18 "cards" of the areas of global Megatrends (each 5–10 pages long) to prioritize from, a document with 1-paragraph summaries of all 18 cards, and an explanation of the context of the project and the design of the expert Delphi.

The question 1 asked *"Choose exactly 6 areas that will, in your opinion, have the greatest impact on the quality of life in Czechia in the next decades and, therefore, public funding should be preferentially allocated to understanding them and addressing them."* Participants were able to tick exactly 6 areas out of 18. This question was posed in order to control for the effect of participants' own values on their ability to forecast group priorities in subsequent research.

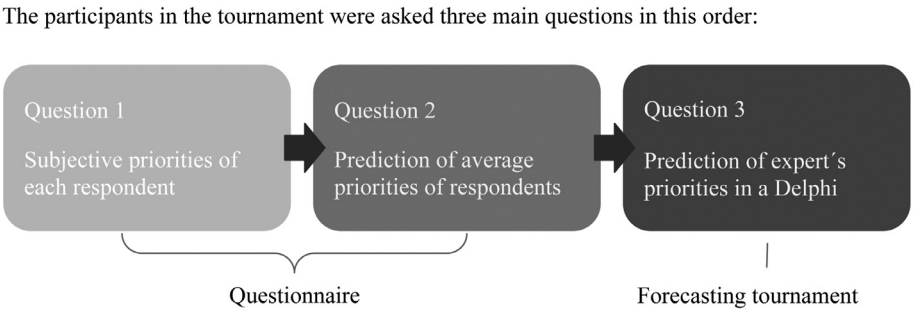

Figure 1. Sequence of the questions.

[57] České priority, "Forecasting project OPTIONS," www.predikce.org (accessed November 2021).

Question 2, which followed, asked *"Which 6 areas, do you estimate, will be selected by the highest number of participants in Question 1 in this questionnaire? The collective score will be derived from a ranking list based on how many times the given area was selected. Again choose exactly 6 areas."* This question was asked in order to better understand the base-rate of the ability of individual participants to predict the consensus of a group regarding these megatrends, before they start participating in the forecasting tournament. We saw this question as creating participants' "track-record" of predicting an opinion of a group regarding these topics.[58] After having answered, participants were allowed to enter the platform.

Forecasting tournament

On the forecasting platform, participants were asked the Question 3 - *"Which of the following 18 areas will rank in the first 6 places of the ranking list compiled on the basis of scores given by experts in the FUTURE-PRO project?"* Answering this question was voluntary in order to limit inputs from participants who would just make uninformed guesses and the participants were asked to write comments and update their predictions as frequently as desired. It was not specified what should be the content of the comments.

Participants had to distribute probabilities of each area being in the TOP 6, which meant distributing the total of 600% between 18 areas. A few participants were confused by this logic, while an explanation had been provided in the first days. The participants were financially incentivized to provide better predictions, as it was announced that after resolution, we would randomly draw 15 participants from all the participants with the above-average Brier score from this question (i.e. from the top 50% of participants), who would receive a voucher in the amount of 1,500 CZK (70 USD). The question was open on a platform for 12 days and the Brier score was calculated each day of the tournament from the Brier scores of each of the 18 areas.

129 participants provided at least one valid prediction. The average participant self-reported spending 95 minutes working on this question. A total of 196 comments were collected, majority of which were phrases such as "first guess" or "updated". The rest of the comments with considerable content could be classified by three main topics - personal opinions on what should be the priorities, comments on how the participant came to their predictions, and the comments on the methodology (both the design of the Delphi and the Forecasting tournament).

To reduce the length of all considerable content, we aimed to select up to 15 norm pages of comments with the biggest informational value. This selection was

[58] This is a standard approach in other crowd consensus mechanisms such as the Surprising popularity. A.M. Rutchick *et al.*, "Does the surprisingly popular method yield accurate crowdsourced predictions?" *Cognitive research: principles and implications* 5, no. 1 (2020): 1–10.

conducted by three independent coders from the project's research team, who ranked all comments on a scale 1–10 according to perceived quality and informational value, which resulted in 13 selected comments. We provided a document with these selected, unedited comments and a two-page summary of all the textual content gathered during the forecasting tournament to the 24 experts before the beginning of the Delphi, which can be found, along with an overview of the 18 cards as well as all the raw comments in the *Extended data* (Kleňha, 2021).[59] After the end of the Delphi (seven weeks later), we resolved the tournament and distributed the rewards.

Ethical approval

Both the questionnaire and the forecasting tournament were ethically approved by the research organization České priority, z. ú. on December 17, 2020. All participants provided consent to use their anonymized data for research purposes by ticking a box to express their agreement during the on-line registration process before the beginning of their participation.

Findings

The combination of the two methods worked as expected and we did not encounter any significant issues during the implementation of these methods. A minor technical limitation was that we could not make the predictions submittable if the sum of all the probabilities was not exactly 600%. This should be, however, an easily solvable problem for future applications. We also found that since most global megatrends are naturally interconnected, it is not always obvious which particular problem is categorized under which area. This can introduce noise to the prioritization of participants who do not allocate enough time to reading all the provided content and prioritize only by the names or short annotation of the areas.

In addition, we investigated the impact of using a forecasting tournament on the ability of participants to predict a group consensus. For this analysis, we selected only those areas that were among the top six priorities by both the aggregate of personal opinions of participants (Question 1) and the results of the Delphi. With this selection, we aimed to limit possible bias in the results of the analysis that could exist if the experts in Delphi choose some of the areas inadequately due to limitations of the Delphi method.

The data that support the findings can be found in the *Underlying data* (Kleňha, 2021). The names of participants are replaced by anonymous numerical identifiers to protect personal data.[60]

---

[59] J. Kleňha, "Improving National Strategic Foresight," *OSF*, October 27, 2021. http://doi.org/10.17605/OSF.IO/94SVE.

[60] Ibid.

Four out of six possible areas were among the TOP 6 in both rankings, namely *Education, Digitalization, Innovation (Science)* and *Environment*. All rankings are available in the FUTURE-PRO final report.[61] For the analysis of predictions in these areas, we used the responses to questions 1 and 2 and observed two aspects:

A. Group accuracy - whether the responses to question 3 were more accurate than the responses to question 2, suggesting that the forecasting tournament increased the group's ability to predict this area to be among the TOP 6 priorities in either of the rankings. The results are visualized in Figure 2.

B. Individual accuracy - whether there were more participants who did not select this area in question 2 and then correctly selected it in question 3 (updated in the right direction) than those who did the opposite (updated in the wrong direction), suggesting that, on average, participation in a forecasting tournament increased the ability of participants to predict an opinion of a group. The results are visualized in Figure 3.

Interpretation

Group accuracy

In all four areas, the average group opinion distilled from a forecasting tournament was more accurate than the average group opinion when a yes/no questionnaire

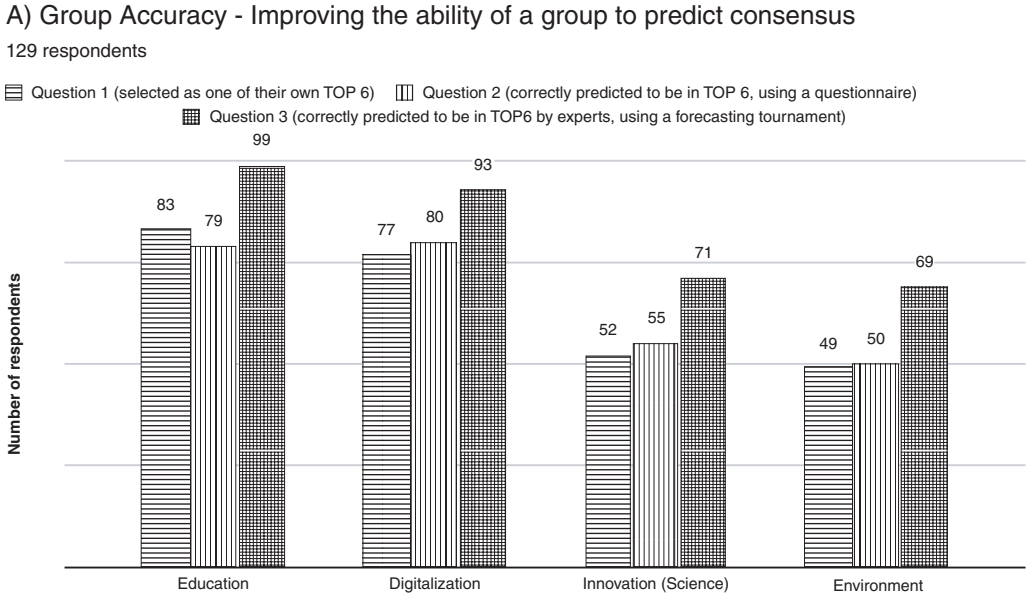

Figure 2. A) Group accuracy – improving the ability of a group to predict consensus.

[61] České priority, "FUTURE-PRO."

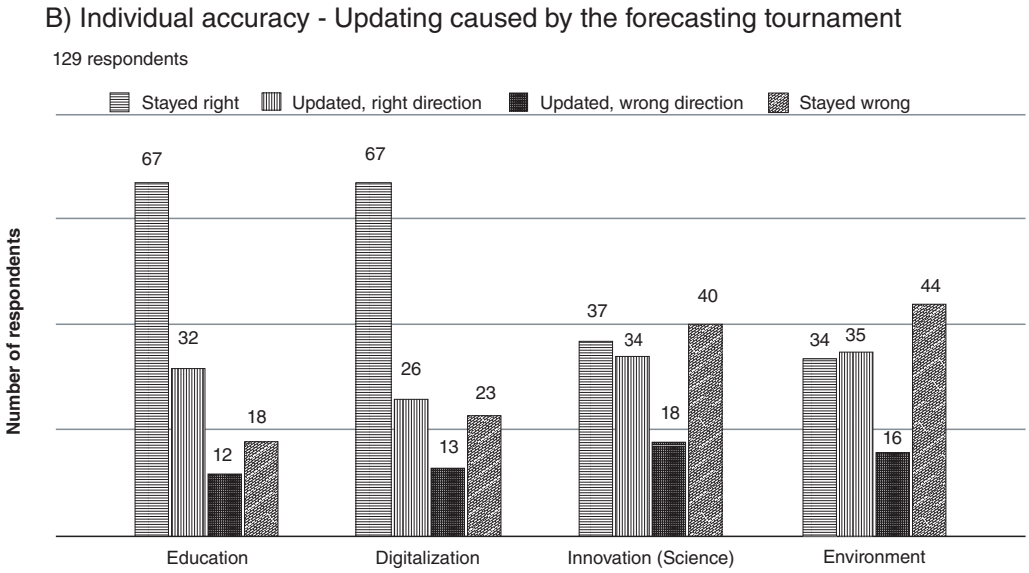

Figure 3. B) Individual accuracy – updating caused by the forecasting tournament.

was used, as expected. Moreover, the forecasting tournament seems to have helped participants, on average, to disregard their personal opinions and values more strongly in favor of accuracy, but more evidence is needed.

Individual accuracy

On average across these four areas, 56.4% of respondents updated their opinion relative to their prior prediction of a group consensus. Among those who did, 2.2x more people updated in the right direction than those updating in the wrong direction. This finding is in agreement with the hypothesis that forecasting tournaments can effectively reduce bias and noise by, on average, improving the individual ability to correctly predict an outcome, in this case a future opinion of experts. Among the participants who updated in either direction, their personal values played a minor role (average Pearson correlation -0.03 across the four areas).

Limitations

During the time between the Forecasting tournament and the Delphi study, the name of the card "Innovation" had to be changed to "Science". Even though both meaning and the content of the card stayed largely the same, this may have introduced noise to the results as the two words have somewhat different connotations and especially the respondents who did not read the content of the card might have prioritized the "Innovation" card differently than they would have if it was labeled "Science".

The respondents spent considerably (up to one order of magnitude) less time on answering questions 1 and 2 than question 3. It is an important feature of

forecasting tournaments that people are motivated to spend a lot more time conducting their own research, writing comments and updating their responses, but we are aware that this discrepancy may have been further amplified by the setup of the experiment. Questions 1 and 2 were part of a mandatory questionnaire, which did not let participants enter the forecasting platform before answering it. This was the best possible setup, as we needed participants to answer questions 1 and 2 before seeing (and being influenced by) others' comments and predictions (question 3) on the platform. We had published the questionnaire three days before question 3 was published on the platform and gave notice to all participants even prior to uploading the questionnaire, so that they could plan their time accordingly, but it still may have been a limitation.

We tried to make the amount of information the participants in a forecasting tournament had about other participants similar to the amount of information they knew about the future experts in Delphi (e.g. they knew the distribution of expertise in both groups, but not the identities of neither group members), but possibly imperfectly. In the analysis of the results, we have not measured the strengths of the effects of individual aspects of a forecasting tournament that are not present in a simpler yes/no prioritization questionnaire (mainly the aspects of answering in probabilities, group information sharing and the possibility of updating), which would be a relevant question for future research.

Conclusion

The combination of Delphi and a Forecasting tournament is a both theoretically and practically feasible design for strategic foresight studies. The findings also suggest that a forecasting tournament increases both the group's and the individual's predictive accuracy relative to a survey or a questionnaire, but more evidence is needed to test this hypothesis. It is possible that the distribution of priorities aggregated from questions 1, 2 or 3 will be retrospectively considered to have been more accurate than the distribution that came from the expert Delphi study. This would be an interesting finding, but it will require re-evaluation on the same topic in the next 3–10 years.

Implications for the study of international relations

International relations and geopolitics are the fields of study where future developments are often the result of a large number of complex processes that are highly difficult to predict using statistical methods. Deliberative methods such as Forecasting tournaments or a Delphi method can, however, be effectively used to provide valuable inputs. The design described in this article could help to further strengthen the ability of nations to formulate more robust foreign policies, but it could be also used by institutions that need to understand, for example, the precise likelihoods of different scenarios of international conflicts.

Directions for further research

Further research could focus on implementing and validating more of the specific aspects of Reciprocal Scoring. It could also explore, by which specific aspects (such as raising the awareness of usual biases of experts or providing minority views or contrarian views) can the written inputs collected during the forecasting tournament reliably serve as a useful source of information for the experts in the subsequent Delphi. Another possible direction is the combination of this methodological approach with the Surprising popularity mechanism, as it may increase the ability to amplify priorities about which the majority of experts in Delphi is wrong.

Overall, improving the methods of strategic foresight seems to be a promising direction, in which more research and experimentation could be potentially very impactful. Increasing the capacity for high-quality foresight should be among top priorities of the national governments and international organizations that understand the importance of making decisions based on robust predictions of future developments.

Data availability

Underlying data

OSF: Improving National Strategic Foresight. https://doi.org/10.17605/osf.io/94sve (Kleňha, 2021).

This project contains the following underlying data:

Data EN.xlsx (the dataset includes participants' responses to the questions [responses to question 1 are listed as "MTPP_Preference", question 2 as "MTPP_Prediction" and question 3 as "MTPR"] and participants' demographic data and questionnaire responses).

Extended data

OSF: Improving National Strategic Foresight. https://doi.org/10.17605/osf.io/94sve (Kleňha, 2021).

This project contains the following extended data:

Forecasting tournament output as provided to the experts in Delphi.pdf

Questions EN.pdf

FUTURE-PRO Methodology.pdf

FUTURE-PRO Full report.pdf

Comments CS, EN.xlsx

Overview of 18 cards EN.pdf

Acknowledgement

I wish to extend my gratitude to the team of the FUTURE-PRO project, especially to Kateřina Jiřinová, Ladislav Frühauf, Marek Havrda and Filip Šourek, as they were essential for the success of the application of this methodological approach. I also want to thank the team behind the forecasting platform Cultivate Labs, who provided us with the software platform, and to the team of the forecasting project OPTIONS.