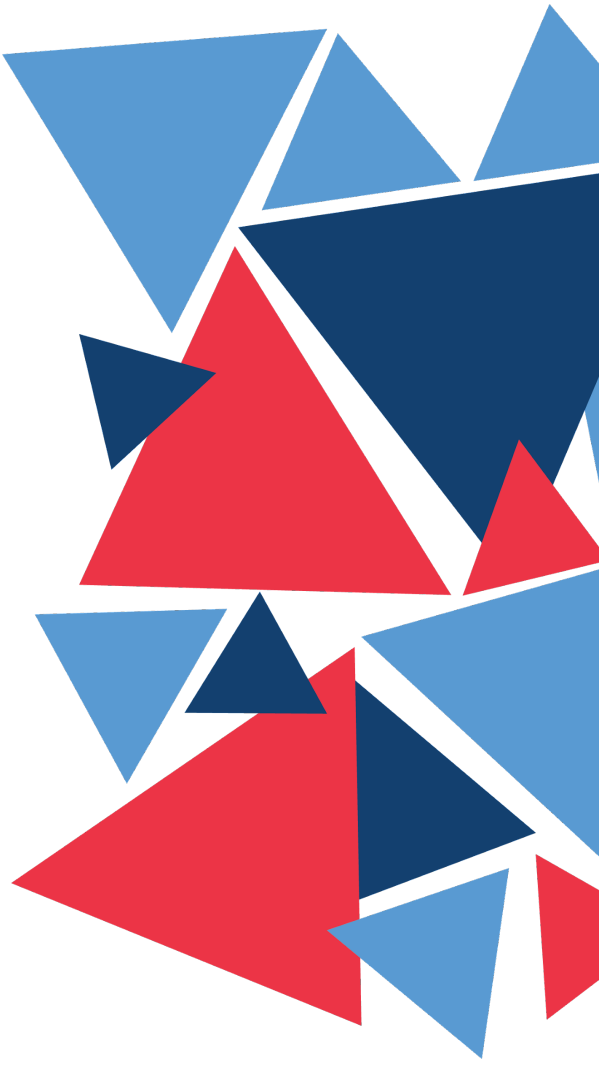
ČESKÉ PRIORITY

Forecasting for Policy

Findings from a Policy-Focused Forecasting Tournament in the Czech Republic



*The **FORPOL (Forecasting for Policy)** tournament benefited from a number of inputs and peer review comments by established representatives of forecasting-related organisations. We would like to thank all those who generously contributed with their time and expertise.*

***Czech Priorities** is a non-profit research organisation based in Prague, the Czech Republic. It was established as a spin-off from the local Effective Altruism movement in 2018.*

Written by Dominik Hajduk, Jan Kleňha, Tereza Majerová and Pavel Hanosek from Czech Priorities, with input from Ryan Beck and Sylvain Chevalier from Metaculus.

Version 1
September 2023

Contact: info@ceskepriority.cz

Table of Contents

Executive Summary

Project outline

Forecasting for Policy (FORPOL) is a project by Czech Priorities, an NGO, in cooperation with Metaculus, undertaken thanks to support from the Effective Altruism Infrastructure Fund. It represents a continuation of our previous work done in this area, most notably the OPTIONS forecasting tournament (2021).

The FORPOL tournament lasted for 6 months. Every three weeks, we presented our forecasters with two new "challenges" - batches of 2-4 policy-relevant forecasting questions that were developed in cooperation with various public institutions from the Czech Republic and Slovakia. The tournament took place on a private subdomain of Metaculus, with the results and commentary then made available to the policymakers as inputs to their decision-making.

Institutions & topics

The majority of challenges were developed directly with individual departments at various Czech line ministries. When planning our outreach, we paid special attention to topics which have thus far been on the periphery of interest in forecasting research - education, digitalization, or social affairs. This was to both expand the scope of current knowledge, and potentially serve as a proof of concept for further work based on identified lessons.

The report proceeds in two parts. First, it discusses the policy-facing aspects of the project, including a description of the process through which we collaborated with public sector institutions. This section also contains notes on what proved effective in communicating both the premise and outputs of forecasting to policy officers and other decision makers. Second, it presents a structured view of the tournament itself, along with some of its specifics such as the use of "unresolvable" questions, and its participants. Some qualitative data is then presented on the forecasts and their characteristics.

Lessons learned - Policy

- **Diversify your foresight portfolio.** See if you can build synergies with other foresight methods. Especially in the EU, policymakers are increasingly becoming aware of foresight (here due to EU Commission outputs). This can help with both initial communication of forecasting and scaling up the partners' interest and effort.
- **Be prepared to be in the driver's seat**. While public institutions might be largely supportive of the idea of forecasting, their ability to closely cooperate for the whole duration of the several months long forecasting tournament is limited. Keep in mind that their primary function is usually not to experiment and discuss forecasting questions and findings.

- **Don't get locked in.** Even if an institution is receptive to forecasting, you may eventually find out that developing feasible questions for their topics of interest is not possible (such as due to data availability, time horizons, etc.). In this case do not feel obliged to submit sub-optimal questions to forecasters. You are the partner with awareness of how forecasting inputs should look.
- **Find the sweet spot.** Policymakers will not want to include probabilistic forecasts in only one chapter in a policy document of many. It would look inconsistent. Aim for policy issues which are likely to have their own standalone discussions and outputs, where forecasting can really pop.
- **Looking ahead.** We found the greatest interest from policymakers in questions of a 7-12 month time horizon, although long-term questions (of 10+ years) were also in demand. Breaking down these long horizons into shorter, repeat forecasts can be of interest.
- **Give them something to think about.** We received very positive feedback on including forecaster rationales and other contextualising information in supplemental materials provided along with pure probabilistic information. Try to aim for anywhere between 3-10 pages for a handful of questions.

Lessons learned - Forecasting tournaments

- **Help them help you.** Scoring rules determine the feedback that forecasters get on their predictions. They need to properly understand how scores are calculated and what they mean so that it serves to inform and motivate them.
- **Mind the gap.** There are numerous factors that will make significant drop-off inevitable in forecasting tournaments (cognitive and time demands, primarily online activity, etc.). Keep this in mind when planning your forecaster recruitment strategy and goals.
- **Variety is the spice of life.** Forecasters strongly favoured a wide range of topics covered in the tournament. There's a need to strike a balance between greater diversity and the greater time investment (for research) demanded by it.
- **Can't win them all.** At the start, there are three important objectives: improving public understanding and acceptance of forecasting; identifying and developing top forecasters (and generally keeping forecasters engaged); and crowdsourcing forecasts useful for public policy. At various times these objectives may temporarily clash. Know which is your priority.
- **Cashing in?** Financial incentives are not a perfect or singular solution for motivating forecasters. Further work needs to be done on identifying the best ways to incorporate them - tournament prize pools should not be uncritically considered a perfect solution.
- **Rationales don't compete.** We offered additional rewards for well thought-out rationales. While this improved the base quality of contributions, we did not observe explicit competition between forecasters in writing outstanding rationales.

Introduction

This study consists of two main chapters followed by a discussion of the data and holistic reflections on the project. In **Chapter I: Cultivating the demand**, we discuss our findings from working with policymakers on asking the right questions and then using forecasts in their work. In **Chapter II: Harnessing the supply**, we share our experience with running a forecasting tournament to answer these public policy questions. The report closes with implications drawn for further work in this research field.

Judgmental forecasting is an essential tool in decision-making, as it allows actors to anticipate and plan for future events, trends, and opportunities. However, the accuracy of forecasts is often limited by a variety of factors, including incomplete data, unexpected events, and human biases. Forecasting tournaments are a method which aims to address these limitations in several ways, including most notably by filtering out the noise which may otherwise significantly affect individual forecasts (Satopää et al., 2021).

Despite research indicating both the accuracy and other benefits of the use of crowdsourced forecasting mechanisms[1] (Tetlock et al., 2017; Stastny and Lehner, 2018) they remain woefully underutilised - especially in public administration. There are only a small number of publicly discussed experiments with the use of forecasting tournaments to inform government decision-making. Moreover, these tend to be contained to a small set of (anglophone) countries, which can raise questions of general applicability in other national/institutional contexts.

A similar issue arises from the primary use to-date of the geopolitical/foreign policy domains in projects. While it is a natural extension of the research on expert judgement in these areas, this focus can lead to the impression that benefits derived from the use of forecasting tournaments are limited to foreign policy. However, there is no fundamental characteristic of judgmental forecasts and/or their aggregation using tournaments which would make this the case. Working with base rates, reducing bias and noise, or even advanced techniques such as extremization are in and of themselves subject-neutral. In fact, given that policymakers in "civilian" domains do not have to claim exceptional knowledge and predictive capabilities by virtue of their access to confidential or classified information, they may be more open to the use of such tools in the first place.

Therefore, we believe that one avenue by which forecasting might gain further traction in influencing government policy is by creating further diverse examples of the value derived from forecasting and its use-cases. To achieve this, we ran a six-month long forecasting tournament in the Czech Republic, in which almost all questions posed were designed *with* and *for* the

[1] We understand such mechanisms to mean both forecasting tournaments and prediction markets, though the latter was not a subject of the presently discussed project.

public sector. In this way, we could look to increase the ownership of both those specific forecasting outputs as well as a case study for the ownership of forecasting tournaments as a process by public sector organisations.

We believe there are three fundamental reasons for why the research team at Czech Priorities was uniquely positioned to undertake this project. Firstly, we already trialled the use of public forecasting tournaments in the Czech Republic with our previous project "OPTIONS". Subsequently, we have not only taken on the mantle of popularising the method domestically, but have also already identified some of the premier forecasters in the country, and formed a dedicated team out of them, which we could then show to our potential partners in the public sector. Secondly, within the public sector itself, our other previous projects on informed governance have given us access to senior policymakers and a reputation within the organisations themselves, enabling us to touch on a broad range of topics while still having a tangible link to the responsible policymaking institution. Finally, in light of the previous two, we believe that the Czech Republic is fundamentally a good fit for a project of this type, as the comparatively low barriers to change on an administrative level can facilitate getting "quick wins," from which forecasting can snowball.

To run the tournament, we partnered with Metaculus, a premier forecasting tournament website, which helped us to create and then allowed us to host a separate subdomain, translated into Czech language, with all the features available to other Metaculus users.

Before we launched the tournament, we spent several months developing a project design document and consulting this with numerous key stakeholders and experts in the international forecasting community. This project design document contained a detailed discussion of key aspects of the planned tournament, such as

- the contexts of "supply" and "demand" sides,
- our theory of change,
- key design decisions and their alternatives,
- original design of cooperation with partners.

Chapter I: Cultivating the Demand

The fundamental premise of FORPOL has been the generation of forecasting questions *with* and *for* public sector, policy-relevant institutions. In this section, we describe the approach we took to secure buy-in by relevant actors. Based on these experiences, we outline the changes we made to our own format over the course of the project, and offer some further recommendations. This is informed by a number of themes which we found to be repeated across interactions when presenting forecasting.

We then proceed by providing a closer description of the institutions we partnered with and policy domains of interest. Finally, we select a few partnerships which we describe in closer detail as case studies on the cooperation in policy-relevant forecasting: as successes, as experiments, and as failure modes.

Collaborative question development

At the outset of the process, our team established a contact database working off previous evidence-based projects undertaken by Czech Priorities. Following this mapping, we reached a long-list of approximately 300 potential points of contact interested in future-oriented thinking. From this, we derived 37 prospective contacts who were expected to have both interest in forecasting and significant sway within their respective institutions to be able to commit (human) resources to the exercise. Approximately half of the institutions represented in this sample had two or more points of contact identified.

Subsequently, we ranked this shortlist based on expected interest in forecasting, potential ongoing collaboration in other projects, and - if relevant - time-sensitivity of the forecasting questions of interest to the given contact/institution. We then reached out to the contacts in descending order, working in batches. Between June and August 2022, we reached out with a formal offer of up to 5 crowd-sourced forecasts in the coming months free of charge to 18 Czech public or non-profit institutions. The partnership was subsequently realised with 8 of these institutions (for a 44% success rate).

In the end, we partnered with 15 institutions - the rest being made up of representatives from e.g. ministerial departments, with whom we already had working relationships (mostly unrelated to forecasting), who showed interest in this as well (with quite a similar adoption rate). As a result, we provided an average of 2.75 on-demand forecasts across 16 reports.

Initial contact

We systematically utilized our existing connections, usually to ask for reference or connection with the most relevant person at the institution. In cases without available connections, we approached the selected policymakers directly. In the initial email conversation, we included a one-pager on the tournament as well as on forecasting more broadly (included external links).

First joint meeting / call

If our identified points of contact responded to our initial e-mails positively (over 70% of cases), we arranged a meeting or a call with them to provide more information and address their unique concerns or use-cases. During these calls, we used a standardised set of slides, first briefly introducing foresight as an input to policy making before bringing attention to judgmental forecasting and the FORPOL project. The presentations, however, had one notable change: in each presentation, we made sure to already show examples of how a forecasting question might look like **for that particular policy domain.**

This was especially helpful as one of the recurring themes in our discussions with policymakers was the difficulty in internalising the types of questions which can be forecasted. More specifically, in most cases (with the main exception being contacts with a strong background in quantitative thinking) the first instinct of policymakers was to pose "How" and "What" questions - "How best can we achieve X" or "What should be done about Y." As long as they were only shown "generic" forecasting questions - i.e. a screenshot of the Metaculus front page - it was difficult for them to imagine how forecasting questions might look in their domain.

Although it happened only rarely that some of our "early suggestions" were taken up into the tournament itself, they had the value of demonstrating how the subject matter with which those individual policymakers were tasked with can be transformed into a forecasting question. As our model relied on one-on-one contact, the option of a broader, open-ended strategy to tackling these challenges was not in line with our project.

From a wider perspective, however, the fact that this was an issue which we encountered across departments and with diverse policymakers hints at two complementary possible avenues for further work which would also benefit the take-up of forecasting for public decision-making.

First, it may prove useful to **create pressure for public organisations to be more explicit about indicators which are of note to them**. Crucially, this means not only indicators which their policies aim to target (or their proxies), but also indicators for factors which may exert some effects (positive or negative) on those. In our experience, even moving towards thinking about monitoring possible enabling conditions or constraining factors was seen as unnecessary for some.

Second is the even more ambitious task of communicating clearly the necessity for policymakers themselves to start (not only) the forecasting exercise with a view of the potential trade-offs of their policies in mind. While it would seem obvious that policymakers are to be aware of the key controversies/decisions to be made[2] in their respective files, on

[2] Such as the exact form a policy should take place, deadlines, etc. For example, a random recent RIA in the Czech Republic notes among one of the decisions that need to be made the scope of institutions that

numerous occasions we have found that there were difficulties with mapping them onto forecastable metrics. **Only in this way, questions about the "goodness" of policies - which will always be on their minds - can be meaningfully operationalized in forecasting tournaments to leverage their unique strengths.** Nominally, this should not be a significant obstacle - in many countries, the notion of Regulatory Impact Assessments or Statements (RIAs/RISs) is long established. However, their development is often disconnected from the early phases of the policy process. Helping to bridge this distance would create further opportunities for forecasting in public administration - ones that would, in our experience, be welcomed by the policymakers themselves.

Finally, we often encountered the demand and the actual usefulness of the use of **other foresight methods**. It proved useful to have a good understanding of the landscape of other available foresight methods, some of them reasonably suggested to the policymakers by international bodies such as the European Commission, to which they should be responsive.
We had the benefit of having previously developed robust foresight expertise, which turned out very helpful in the initial discussions. Unsurprisingly, forecasting is viewed with less suspicion if those offering it are responsive to the unique circumstances of individual organisations, recognizing that certain types of questions are more reliably addressed with other methods. Moreover, the precise understanding of other qualitative foresight methods (horizon scanning, visioning, scenario planning, backcasting etc.) provides the ability to suggest a **combination of forecasting with other method(s)**, which might improve the quality and the usefulness of the outputs. Furthermore, this can then be more readily accepted by those who already trust or know of the other methods.

Asynchronous collaboration

At the end of the first joint call, we shared access to a collaborative workspace where blank documents were available with a structure aimed at eliciting potential question topics and formulations asynchronously, in advance of a second meeting.

We found this to be necessary for practical reasons. Even though they had received a one-pager on forecasting, partners often had many (usually straightforward, though in general varied) questions during the calls, and only after answering those were they able to start thinking about the questions they might need answered.

To guide partners towards realistic and forecastable questions, we used tables - one for each overarching policy goal (*"What do we want to achieve?"* - i.e. "A higher-quality school counselling service") - with four columns to be filled in. These columns were as follows:

- **Question topic** - the first column was populated usually with the very rough question ideas emerging from the first call, such as "Who should be...", "What is worse:..." or

would fall under new regulation - should it follow the list used by similar regulations, or be modified in some way - to be either more or less permissive?

"How to...". By themselves, these are often not operationalizable questions for a forecasting tournament, yet their inclusion helps both reassure the partners that their topics of interest can be further explored, and engage them to familiarize themselves with the logic of the following columns. If appropriate, further rows were then started off with ideas of our own and/or left blank to be filled in later.

- **Goals of the question** - *"What is needed to achieve the preceding."* For example, if the first column identified "Reduction of school deferrals," here one might place "Showing that directed action is necessary, i.e. the problem will not solve itself on its own."
- **Selection of indicator** - *"How can we know?"* This column represents an attempt to structure the transition from the qualitative mode of thinking many policy officers encountered were used to to a more quantitative streak.[3] Importantly, this column was not to be filled with the intention of already identifying a specific indicator (though it was not discouraged). Rather, the focus was kept on trying to map the previous answer onto something that is potentially measurable or falsifiable.
- **Question** - Here, the previous answers were all combined into a single forecasting question. Though this stage was often left to us, given the specifics of forecasting questions (setting time/answer scales, selecting the resolution source, etc.), it was still practical to have the column visible in the document. Being able to see and track the thought process matched to a specific forecasting question reassured our partners and acted as a good record-keeping exercise.

The following table offers some examples of what such an approach may look like at the end:

Question topic	Goals of the question	Selection of indicator	Question
How can we support informal and lifelong learning?	*Can we rely on the network of public libraries as a key pillar in informal learning strategies?*	*Existings statistics track several different things: books loaned, number of libraries, registered library cards - each showing part of the picture, but blind to other aspects.*	***How many books will be loaned nation-wide in Czech public libraries in 2025?***
How can we support informal and lifelong learning?	*Is Czech society on the path to greater acceptance of alternative qualifications?*	*Non-academic paths in some areas will only truly be seen as alternatives if institutions demonstrate non-prejudice*	***What percentage of Czech universities will award credits for informal learning certificates in the academic year 2025/2026?***

[3] Quantitativeness need not imply wholly numerical data - binary indicators were, of course, also used

		themselves.	
How can we support informal and lifelong learning?	*Will demand for job retraining rise?*	*Demand for retraining will rise if there is confidence in its effectiveness, and there are others showing that it is meaningful.*	***Will the share of retraining costs on the total active labour market policy costs exceed 10 % in any one year before 2026?***

Second joint call

At this point, the content of the discussions differed noticeably. Generally, we spent 1-2 hours asking additional contextual questions, presenting our suggested forecasting questions and validating whether these forecasting questions met both their needs and expectations and the items on our "checklist" of a proper forecasting question.

The most important part of the discussions at this stage turned out to be the intended future use of the predictions. With hindsight, it seems to us even more important to get thoughtful responses to questions such as

- *"What will you do, when the prediction falls under/above a certain threshold?"*
- *"Which of your ongoing analytical work will these predictions feed into?"*
- *"How and with whom exactly will you communicate the results?"*
- *"What could prevent the prediction from being used as planned?"*

Communication during the forecasting period

Curiously, we did not encounter significant interest in closely examining the mechanism by which forecasts were made. While we offered to showcase the Metaculus subdomain to interested partners, in the end, none actually requested a demonstration or access to the platform. There could be several reasons for this: from the fact that the attention of (senior) bureaucrats is a limited resource, to want of an (epistemic) distance from the process and its outcomes.

Of course, it remains an open question whether to consider this phenomenon inherently good or bad. For example, one might take the view that this distance allows for a quicker and more efficient scaling of forecasting in the public sector, which then directly links to the probability of a critical decision being made with forecasting inputs in mind.

Based on our discussions with other forecasting-involved organisations throughout the project, however, we cannot ignore another line of thought. In this view, only embedding the policymakers into the forecasting process itself will maximise their willingness to use its probabilistic outputs. Furthermore, expanding their understanding of *why* such cognitive

models would improve their work might be a necessary factor in securing the sustainability of the use of forecasting across their respective institutions. In general, it would seem useful to pay close attention to the main constraints on partners' participation - and if lack of time/attention is not the most significant one, to maintain open communication - i.e. by sharing topics of other ongoing forecasts if they may be of interest - during this period.

Reporting back with the forecasts

Following the closure of the forecasting window for each individual challenge, a short report summarising information from the forecasts was prepared and shared with the partnered institution. As part of this communication, we also offered both a) a more comprehensive debrief call, and b) a list of other organisations which may find the information contained in the report of interest, in view of facilitating dissemination and maximising the value to decision makers.

The reports themselves ranged based on the number of questions and their complexity. However, as an internal benchmark, we strove to deliver approximately 300-500 words of contextualization (previous development, main influencing factors, etc.) and noteworthy quotes and justifications from forecasters to complement the numerical data derived from the Metaculus forecasts for each question.

Across partnered institutions, an almost universal unprompted feedback was the gratefulness for this contextual written information - our contacts rated it very highly, and considered it critical for using the forecasting outputs in their further work, but also often could not point to any direct causality linking the forecasts to a discrete policy or its outcomes. This was usually because a practical implementation did not yet happen or the impact was done through unofficial consultations with leadership or other stakeholders, where forecasts may compete for impact with other sources, with difficulties in assigning individual contributions. Based on this experience, we would **strongly recommend any future projects make sure their outputs provided to decision makers do not omit such sections**.[4] On the other hand, we were mindful when collating these reports of not crossing the line of explicitly advocating for any specific policy course or action, instead letting the identified factors and data (both existing and predicted).

Partnered institutions

Throughout the project, we tried to cover as many underrepresented (in forecasting) policy domains as possible. As a result, FORPOL fielded questions on topics ranging from elementary school deferrals, through the effects of climate change on Czech regions, to the adoption of modern technologies by citizens. Each topic was the subject of a "challenge" - a three-week

[4] This echoes the findings reported by Samotin et al., who state that in their interviews, "Interviewee 25 reported that, in his experience, intelligence consumers struggled to interpret probability assessments without receiving information about the logic driving analysts' judgments, which the ICPM and the ACE program did not *synthesize systematically*" (p. 18 of author preprint, emphasis ours).

window where forecasters could submit forecasts on 2-4 questions on that given topic. Figure 1 below provides an overview of the titles of individual challenges, while Annex I contains also the wordings of individual questions grouped by challenge.

This is not to say, however, that we completely sidestepped established forecasting use-cases (for example, we had epidemiology or macroeconomic challenges). However, we prioritised policy domains where - to the best of our knowledge - there has thus far been little documented use of probabilistic inputs into the policy process.

Subsequently, this combination allowed us to draw some preliminary conclusions while keeping constant the forecaster population and (to the extent possible) national/organisational contexts of the partnered institutions. For example, while working with institutions from the more established domains saw, on average, fewer barriers in explaining the potential use cases and demonstrating the use of forecasting, this was not always the case. Instead, some of the institutions from unorthodox policy domains demonstrated greater interest in and willingness to use forecasting than those in the economic/geopolitical domains. However, due to our relatively small sample size, it would be preemptive to draw any substantial conclusions from this observation other than the fact that one should not discount the interest in forecasting in unorthodox domains.

Based on these experiences, however, we can attest to the influence of "champions," i.e. senior officials who push for the use of forecasting in their respective departments, which is often highlighted in the literature in traditional forecasting domains. For example, the belief in the ability of forecasting to bring valuable insights led to repeat interactions with the Ministry of Education. Therefore, it is one of our key findings that such **champions can be found and cultivated throughout government.**

It is important to note, however, that relying on a few key relationships can have negative effects even in a short timeframe. For example, a small number of our contacts suggested leads for forecasting topics which after some further research were found to be sub-optimal. However, given the invested time and effort on the side of the public-sector contact, it proved difficult in some cases to reconsider/drop the topic. Therefore, **while pursuing a wide-net strategy was more demanding than focusing on established contacts, we would recommend further projects continue with this approach.**

In total, we partnered with 15 unique *primary organisations*, i.e. those organisations which directly helped shape, formulate, and resolve the questions. Of these, 10 were either Ministries or government agencies, with repeat interactions in the role of primary organisation by both the Ministry of Education and Ministry of Regional Development. Further Ministries included those of Health, Finance, Industry, Labour, Justice, and the Ministry of the Interior. Additionally, we worked with the Cabinet Office and the Technology Agency of the Czech Republic, the main public R&D&I funder.

The three remaining primary organisations were a think-tank, a research organisation, and a policy-focused NGO, respectively. In these cases, we collaborated with the organisations due to either their positioning in sharing the forecast beyond our own reach, or due to their access to data which could be used to resolve questions of interest to government organisations. In all three cases, however, the ultimate goal was to provide policy-relevant information to decision makers, and in two cases (leaving only the public research organisation) a *secondary organisation* with interest in the forecasts was already identified in advance.

Furthermore, we found that on numerous occasions, forecasts originally commissioned by a primary organisation piqued the interest of secondary organisations after the fact, i.e. when results were already available. Unsurprisingly, this tended to happen more often with questions related to extremely high-salience topics (i.e. refugees from Ukraine) or with questions with longer time horizons (and thus with greater potential to impact multiple policy domains). As a result, we would recommend future projects to **discuss the shareability/interoperability of forecasts commissioned by various parties throughout a tournament upfront.**

In sum, of the 18 challenges which were posted on the FORPOL Metaculus domain, all but 2 (the *internal questions*) were developed in cooperation and for the purposes of a specific policy-relevant partner:

1. *Internal questions*
2. Housing
3. Urban development
4. *Internal questions II.*
5. School deferrals
6. Technology take-up
7. Covid-19
8. Politics I.
9. Staffing in schools
10. Public opinion
11. Women in IT
12. European macroeconomics
13. European industry
14. Household debt
15. Lifelong learning
16. R&D&I financing
17. Forensic experts
18. Politics II.

From a conceptual viewpoint, we partnered with institutions to supply them information (in the form of a probabilistic forecast) at all five policy cycle phases.[5] The following table shows the distribution of questions and topics across the cycle:

	Agenda setting	Policy formulation	Decision-making	Policy implementation	Policy evaluation
Internal questions					
Housing		■	■		
Urban development	■				
Internal questions II.					
School deferrals		■	■		
Technology take-up	■	■			
Covid-19			■		
Politics I.	■	■	■		
Staffing in schools	■	■			■
Public opinion	■				■
Women in IT	■				■
European macroeconomics		■	■		
European industry	■	■			
Household debt	■	■			
Lifelong learning	■	■			
R&D&I financing			■		■
Forensic experts				■	
Politics II.	■	■	■		■

[5] As per Howlett and Giest, 2015

This shows a clear tendency for feasible policy-relevant questions within the institutions we interacted with to be prevalent towards the start or end of the policy cycle. Moreover, the inclusions in the "Decision-making" and "Policy implementation" phases were specific in their respective categories by concerning "rolling," administrative levels of policy rather than high-salience initiatives.

In the "Agenda setting" phase, forecasts were often used for the consideration of ground assumptions on which further advocacy work was to be built. Moreover, one NGO which cooperated with the Ministry of Labour on a controversial policy debate saw value in producing forecasts to generate "depoliticised" estimates on which to build further strategic communications and stakeholder interactions. However, numerous departments wanted to use forecasting as an independent review of what they should (or should not) be concerned about.

In the "Policy formulation" phase, partnered institutions were interested in the way and intensity with which they should tackle issues which they have already identified. There is no shortage of opinions about "what ought to be done" in the Czech Republic, although policymakers can often be stranded once they need to consider how to achieve and prioritise those goals. Running conditional questions or offering a second opinion to contrast with their own without the risk of falling into tired intra-institutional arguments was valued highly by our partners.

In the "Decision-making" phase, forecasting served as an input to the process by which competing proposals on what action to take on a particular issue were judged. Here, what should be focused on - in our experience - are situations where an area can be identified in which some piece of information is unavailable and left blank for each decision maker along the chain, as "we will never know for sure." In these situations, even supplying an 80 or 90% confidence interval value can prove a catalyst for meaningful action or deterrent from counterproductive decisions.

We had only one specific case where FORPOL questions significantly related to the "Policy implementation" phase. As noted, this was related to a specific ongoing initiative and the necessity to shape the delivery of its goals over time.

In the "Policy evaluation" phase, questions were no longer concerned primarily with the parameters or intentions of any planned policy, but rather the impacts of existing ones. As can be gleaned from the figure above, no partner was interested solely on this level, and even where it appeared, it was usually secondary. One creative use for forecasting in this dimension was crowdsourcing Fermi estimates for known unknowns - see below.

Unrealized partnerships

We initiated contact with a broader group of potential partners at the outset of the tournament, a number of which we ultimately lost somewhere along the above-described journey.[6] Most commonly, their representatives either excused themselves on the basis of heavy workloads or stopped responding altogether.

Of the potential partners who dropped out during the question development phase, we believe it is possible to generalise the characteristics that sometimes led to this beyond a simple lack of time (though that certainly still remained a factor). Firstly, certain smaller/less formal organisations may be too "agile" in their activities and priorities to properly benefit from forecasting in this way, and as such may not be able to reap its benefits sufficiently. Secondly, interest in crowd-sourced forecasts may rise with acute crises (such as the European gas supply situation in 2022/2023) and wane as those clear up and familiar topics reemerge, where forecasting is - in the partners' view - "no longer necessary." Finally, a third common characteristic that emerged across several partners was the inability of establishing "indicators" (i.e. data points which can be forecasted) sufficiently quickly/straightforwardly - for example, when developing these from broader qualitative goals, such as in urban development or housing policy.

Perhaps interestingly, with greater proximity to political/executive functions, a (declared) sense of powerlessness discouraged participation. For example, some parliamentarians we consulted were worried that they are not in a position to act on anything learned from a forecast (note that these were not backbenchers, and were given examples of both short- and long-term forecasts).

On a broader note, what these unrealized partnerships had in common was that they were almost entirely institutions or individuals with a broader agenda, which could have led to paralysis regarding which topics to prioritise for forecasting, or else lead to less interest in any particular question which could be developed.

As our rules required a minimal proportion of questions to be answered (see below), we were wary of overloading the tournament with too many questions at any one time. However, this created a difficult situation whereby the scheduling of forecasted questions might have contributed to the omission of certain topics. Therefore, our take-away for further projects would be to build relationships with partners which are from the start understood as ongoing, rather than discrete/one-off. If all partners understand the development and posing of questions as a repeated, two-way conversation, this would minimise the "missed opportunities" as questions could be dynamically shuffled around. However, our expectation

[6] There were approximately as many tapped-but-dropped institutions as we had final partners. Even more institutions were informally approached, but i.e. could not identify a good fit for cooperation.

would be that this would also lead to fewer active partners in general, and thus we did not shift to such an approach during this project.

Feedback from partners

After the tournament was over, we reached out to the collaborating primary institutions to gather their feedback on the process and the deliverables vis-a-vis their policymaking. We received responses from just over a half of the institutions - and only one had indicated that they had not yet had the opportunity to derive any benefit from the provided reports. However, the most common type of benefit was the possibility of "trying the method out," so that it may be considered in the event of significant shocks in the policy environment. Closely followed the benefit of having an aggregated second opinion on the questions being considered internally. While one respondent was optimistic about the potential in policy planning, they noted that this is driven primarily by political cycles still.

Though our final sample is quite low, we wanted to consider the extent to which our partners would self-identify with the potential barriers to forecasting use in public administration proposed and identified by Samotin et al. (2022). Two aspects dominated - one in agreement with, one in contrast to the findings from the US. While "bureaucratic politics" - the lack of professional incentives and absence of a central office or guidelines for their use - topped the list in both cases, our partners stressed the lack of available time to dedicate to the cooperation much more than would have been expected from the results of the former study. There, only 8% of respondents aligned themselves with the sentiment that "Busy analysts do not have time to experiment with new platforms," yet for us it was the second most common concern - with partners reporting between 3 and 20 hours spent on our cooperation, with a median of 8.

As for the prepared reports, our partners on the whole commended the concise nature while still extending beyond a simple numerical value, though one mentioned possibility for improvement was allowing partners to discuss the report with the forecasters directly - something that might be unworkable in a mass participant forecasting tournament, but could be (and indeed we plan to further investigate) included in any future projects focusing on smaller groups.

Finally, during the question generation process, we noted how difficulties could arise with operationalizing policy dilemmas into forecasting questions. Thus, we asked partners to identify what time-frames they considered the most relevant for their own work, with knowledge of how the process went and how the outputs could be used (or not). At least half of the respondents identified a resolution time frame of 2 months to 10 years, although **the single most noted interval was that of seven months to a year** - showing that future research may need to be extended beyond the six months of active forecasting within FORPOL to maximise feedback and learning for both partners and forecasters.

Successes

Our two most clear success stories demonstrate the diversity of benefits which can be addressed to institutions when presenting forecasting.

In one of these, forecasters had the task of predicting the outcomes of public opinion surveys ~3 months in advance, predicting the short/medium term development of generalised trust and other key indicators. The crowd prediction on these showed quite remarkable **accuracy**, to the extent that the partnered institution was taken aback and showed interest in incorporating forecasting into some of their workflows in the future.

A second type of benefit concerns the partnered institution, rather than the predictions themselves. Specifically, when developing questions on the topic of R&D funding, it was communicated throughout the cooperation that the partner's interest lay beyond the predictions themselves. They were more interested in **familiarising** themselves with the principles so that they can evaluate if **building internal forecasting capabilities** is something to devote resources to.

Experiments

In one case, we agreed with a partner who developed several "challenges" to experiment in the setting of more complex forecasting questions. Namely, these were conditional questions or, more specifically, counterfactuals of various policy interventions.[7] This was the one notable case where we structured the whole challenge with this in mind, yet a key finding was that a **substantial number of stakeholders were interested** in such exercises.

There were two practical reasons for why this was (for now) a stand-alone attempt. Firstly, we felt comfortable with trialling such an approach due to the fact that the partner in question had committed to several different "challenges," meaning that their understanding and evaluation of forecasting would not be circumscribed by this one experience. Given that one of our main goals was developing an understanding of forecasting as a method among public institutions, this was a key issue for us. Secondly, as the task was forecasting the effects of distinct policy interventions *ceteris paribus*, we would expect forecasters to have to develop a much more detailed view of the factors in play and their interactions to be able to distinguish between the various scenarios. In preparation for this, we arranged for a meeting between interested forecasters and subject matter experts - in this case, education researchers. The practical impact of this were additional costs tied to their time, which would have made additional attempts at this preliminary stage unnecessarily burdensome, especially as forecaster involvement on these questions proved slightly under expectations.

[7] See Annex I, section 5

Failure modes

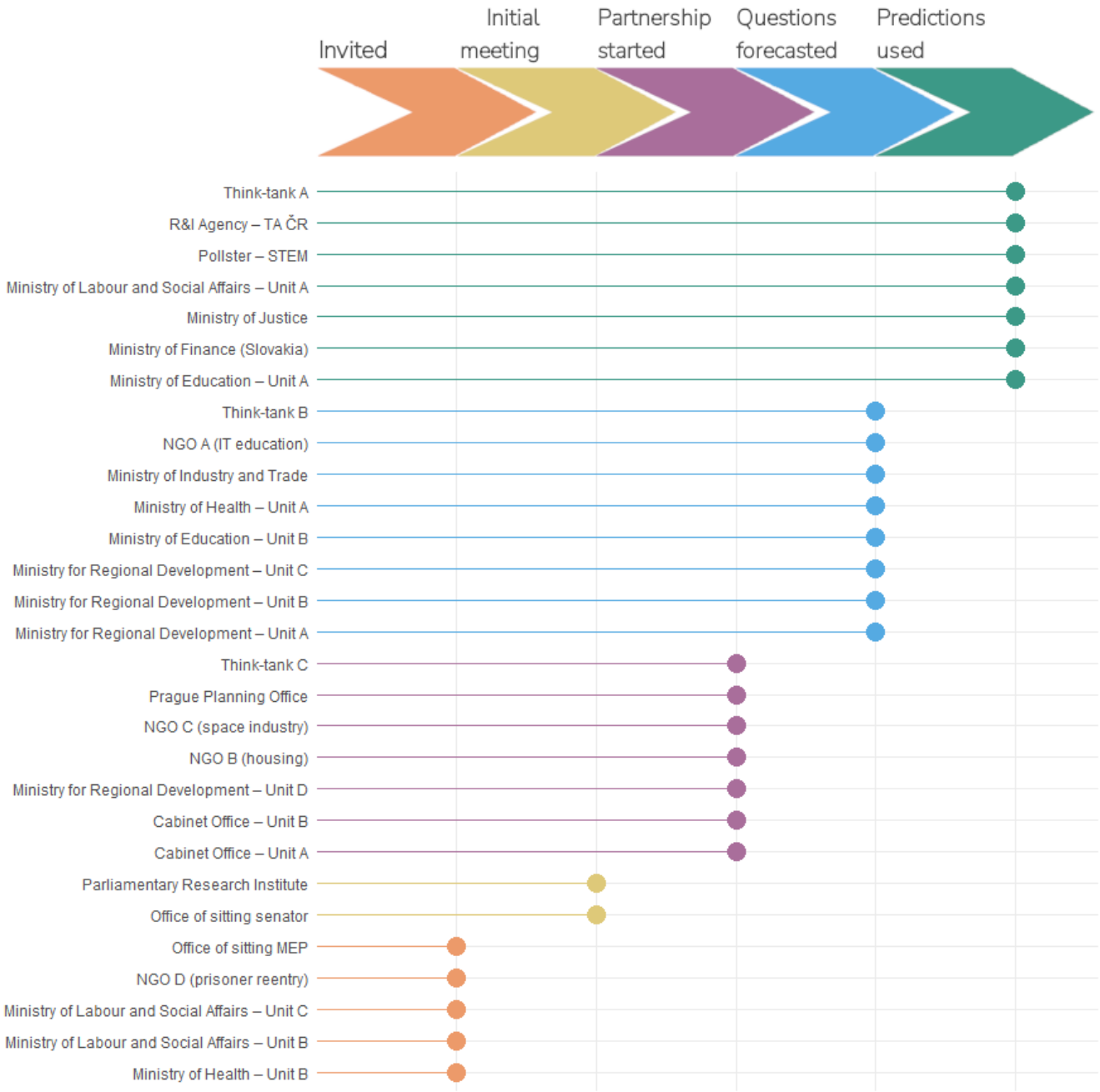
Just like was the case with our successes, even unsuccessful partnerships came in different flavours, which can serve as early indicators for course correction in future projects. These two failure modes can be characterised as tied either to the textual or probabilistic parts of the provided forecasts. Note that in both cases the word "failure" is relative, inasmuch as the partnered institutions may even have expressed the wish to explore forecasting in the future - it is simply that the realised partnership in some way fell short of what they had expected.

For the former, a small number of partners expressed the belief that they would have liked more substantial commentary by forecasters as to the factors impacting their prediction. However, given the topical structure of cases where this arose, we suspect this may be more prevalent in highly specialised topics, where forecasters defaulted to the crowd more often, as they lacked strong beliefs indicating either direction. Moreover, comments tended to be rewarded by peers more often when sharing novel information sources rather than providing commentary/justification of the forecasters' own thought processes. In other words, it's possible that this problem **could be addressed by utilising a different expectation and incentive structure** for forecasters - namely one that would assume participation and reward argumentative contributions more evenly to accuracy contributions.

As for the failure mode of probabilistic information, we don't understand this to mean poor accuracy or similar events. Rather, what is meant are cases where, for example, information was relayed in probabilistic form, and despite individual public servants being enthusiastic about this, it was translated or omitted from the final form of its related policy document. We believe that a cause of this may be situations where forecasts are provided only as a very partial input to a much larger process – for example, if a governmental strategy is updated, and only issues within one aspect of one chapter have been analysed with forecasting. In such cases, directly quoting probabilistic information may distract or else attract unwarranted scrutiny compared to other sections, as it will seem out of the ordinary. Going forward, one way to prevent this would be to **identify and pursue opportunities of smaller government deliverables**, where substantial amounts of these would be contextualised or supported using probabilistic information, so that it seems natural rather than an inconsistency.

Summary

This is the overview of how far we got with the partnered institutions as of September 2023:



Due to the novelty of forecasting tournaments for Czech public institutions and the resources dedicated to developing forecasting questions in the first place, we were not in a position to undertake a comprehensive round of interviews as was attempted in the wake of the Intelligence Community Prediction Market (ICPM) and Aggregative Contingent Estimation (ACE) programmes (Samotin et al., 2022). Our experience seems to confirm, however, their results that "Bureaucratic politics" is the single most frequent obstacle to implementing crowdsourced forecasts in government analysis, though much more so than that study, we ran

into the issue of “Ease of use” - the time required for analysts to get acquainted with the new method and tools.

In other words, we observed that governmental departments as a whole certainly have an appetite for information - the difficult part is getting them to act on it. In light of this, we wonder whether the correct question to be asking going forward might not be *“Will public institutions respond to forecasts?”* but rather *“Are crowdsourced forecasts more likely to be responded to than any other individual piece/source of information?”* It seems that this could be the more appropriate benchmark in the context of which forecasts for policy should be evaluated and promoted.

Any future projects examining such a question would do well to follow what has shown to be the strong points of our approach:

- Regarding finding collaborators in the public sector, keep in mind that these can be found across policy domains, and **champions** for forecasting can be cultivated wherever they may be found. At the same time, however, these champions should not become the sole channels of promoting forecasting - at least in the initial phases - as the project may become “locked-in.”
- As for the produced outputs, we recommend developing and communicating a strategy for the dissemination of forecasts and/or any supplemental materials. Implicit in that is that these supplemental materials (i.e. narrative elements, a written overview of the balance of opinion in forecaster discussions, etc.) should not be considered an afterthought, but an important part of the communication strategy.

As for the thematic focus of the challenges, we can clearly attest to a demand for forecasts even beyond the traditional geopolitical/economic domains most frequently associated with forecasting tournaments.

Relatedly, we found two recurring themes among several different departments, in what we suspect might be a diagnosis not limited to the Czech Republic.

- Policy officers might struggle to access all data available to them as public sector workers due to lack of training, clarity on use restrictions, language barriers, etc.
- Pre-existing predictions and forecasts were done as discrete events, with effectively no way to systematically update them or rerun them with other assumptions - even if there is appetite for that among policy officers and other stakeholders. This can be a good avenue to follow when looking for potential demand for crowdsourced forecasts.

Chapter II: Harnessing the Supply

Tournament outline

Our tournament ran for 6 months, during which questions were posed in 3-week windows. Specifically, every three weeks, two "challenges" of 2-4 questions on related topics were posted on our Metaculus subdomain. Typically, we revealed upcoming questions on Monday noon, with the start of predictions (and the end of predictions on the preceding challenges) at 7pm on Wednesdays. We chose this approach to minimise potential disadvantages stemming from the limited running time of the challenges. With more time in advance to potentially see and think about the questions, the first-mover advantage in scoring could be mitigated somewhat in favour of a wider participation.

For scoring itself, we used the Metaculus points system, which uses a proper scoring rule (expected points are maximised if honest credence is provided). Furthermore, with this system, points are awarded both for absolute (being right) and relative (being more right than other) accuracy of predictions.

Throughout the tournament, we posed 52 questions, with one grouped question of 3 possibilities, for a total of 54 possible predictions - **for clarity's sake, these 54 will be discussed as "questions" below**. In order to increase participation (and thus also increase the value of the predictions for our public sector partners), we set our rules so that only forecasters who responded to at least 75% of the posed questions would be eligible for rewards for their final leaderboard placement. These rewards were structured so that the top 30 placed forecasters at the end of the tournament would receive monetary rewards, ranging from approx. 1400 USD for first place to approx. 120 USD for the 21st-30th place finishers.

Of all questions posed, only two challenges were not created in collaboration with a public sector institution. One of the two challenges with which the tournament launched was populated with simple, discussion-generating questions to ease new entrants into forecasting as an activity. The second non-partnered challenge was a set of two straightforward questions related to political activities on 17 November, a state holiday, intended to increase engagement and stir discussion on the site.

As forecasters had access to a full-feature Metaculus subdomain, they could back up their forecasts with written comments, although this was not mandatory. In OPTIONS - our previous tournament - the provision of rationales was required for each forecast, but we did not find that this had a significantly positive effect on the discourse when evaluating it. Instead, many forecasters often resorted to writing rationales such as "updating" or "based on discussions below," which only diluted those comments which did contain novel pieces of information or arguments. However, as part of the reward structure, we set aside prizes for most valuable

comments, to be awarded within each challenge (so that the complexity or controversiality of a topic would not directly affect the expected value of contributing).

Questions

Types

Of the 54 total questions we posed, only 10 were binary, with the rest being questions with continuous answer ranges. No questions used logarithmic answer ranges, and closed lower/upper boundaries were only used if such options were logically ruled out.

Subject areas

As our goal was to provide value to policymakers throughout the tournament, we had to select policy-relevant questions, ideally actionable and with a sub-year resolution rate. While such questions were not impossible to find (see Annex I for full list), they were often found in quite complex/niche subject areas. This created a significant challenge for us - even if the underlying thought process could be summed up and explained intuitively, the necessities/conventions of forecasting obscured this and could dissuade some forecasters.

For example, the idea “Did Covid-19 just cause a temporary dip in cash use, or have behavioural patterns irreversibly shifted?” became “Will ATM withdrawals account for more than 12% of the total number of credit card transactions in the country in 2022?”

While this in some regards is a natural result of the question decomposition process, we struggled with ways to address it given the structure of our tournament. The ideal way would have been to increase the involvement of forecasters in the question development process (or at least making it more transparent to them). However, if the goal is to develop questions with policymakers in a noncommittal fashion until a suitable question is found, the involvement of a broad community of forecasters might slow the process down. Furthermore, if the questions sought are short-term, the likelihood that only policymakers in executive functions will be interested in them (as they are the only ones who can shape policy on such short notice) increases. Such executive/political policymakers might then be even less interested in a broadly participatory question development process.

In future projects, one way to circumvent the above could be working backwards. This would mean developing with forecasters questions which may be relevant to policy - improving their buy-in and understanding of the questions and their rationales - and only subsequently offering the questions to relevant policy institutions.

Another - though not mutually exclusive - option might be to forecast on policy questions only with smaller, more agile, recognizable and accountable teams of forecasters. This could, we believe, circumvent some of the hesitation in broadening the question definition process.

Framing effect in question details

One phenomenon which we became aware of early in the tournament was the extremely strong framing effect of the details provided with a question. We found that if only a few links to further sources were listed, many of our forecasters would limit themselves to these (as attested by their written comments); however, increasing the number of provided links did not counteract this.

We expected limited forecaster effort spent on information gathering to be one of the tournament's potential failure points. Accordingly, we designed rolling financial rewards for "informative comments" meant to incentivize independent search for and sharing of additional sources. However, while this did lead to some very thoroughly researched comments, **we did not find forecasters competing for these rewards, i.e. by trying to outdo each other's contributions.** As the rewards were quite significant, we think there could be two main reasons for this.

Firstly, there could be a fundamental inconsistency between trying to foster competitiveness in a fundamentally cooperative task (sharing information with other forecasters). A softer version of this claim might instead be developed from something we discovered during exit interviews at the end of the tournament (discussed in greater detail at the end of this chapter). Namely, already when registering, our forecasters had a clear personal rationale for joining the tournament, and this was not very amenable to change from the outside. Forecasters who joined to test their own skills - or simply compete against a friend or colleague - did not reevaluate the "value proposition" of the tournament in light of further incentives.

Secondly, the propensity of forecasters to share a "literature review" might radically decrease simply when seeing another user had already posted a similar list. Irrespective of any potential financial rewards, then, it might seem a more effective use of their time to shift their attention to other questions, or to the development of their own predictions themselves. While seemingly rational, the implicit claim that each individual is capable of objectively assessing when forecasts have reached an "ideal" information saturation seems too strong to us to take at face value.

If neither creating further incentives nor modifying the information baseline seemed to promote grassroots information gathering, what other strategies might be used in future projects?

We would consider trialling a split into information gatherers and prediction producers, as suggested by Alex Lawsen and Nuño Sempere. While their suggestion came as a response to slightly different incentive alignment issues, we think it could elegantly address also the above. For example, splitting the forecasting activity into separate actions, allows for a clearer link to be established between one's spending time and a specific skill. It would also make information gathering an explicit "responsibility" for some users rather than a non-excludable

good dependent on none. Furthermore, it could provide greater direction to the activity, especially to users new to forecasting - (more explicitly) decomposing the process into individual steps might make for an easier onboarding process.

Long-term & unresolvable questions

Already when designing the tournament, we knew that we would like the ability to ask even questions which may not be resolvable in the 6-month span of the tournament. Such unresolvability might stem from two main aspects: either the ground truth would only become known (far) beyond the end of the tournament, or it would never be possible to establish one at all. This is because we believed that expanding the tournament to include such questions would help improve its policy relevance.

While the first type and its relevance to policy might be imagined quite easily (i.e. “By 2050, will the Czech Republic experience floods with estimated total damages in excess of 120 billion in January 2022 prices?”), the second category warrants a closer look. In the tournament, there were two main ways in which we used this approach.

Firstly, it allowed us to coordinate a Fermi estimation exercise for a known unknown - “How many kindergarten and elementary school teachers who taught in 2019 were no longer teaching as of 1 February 2023?” Such data is (regretfully) not collected and stored by the Ministry of Education, despite its obvious value for addressing teacher turnover. Secondly, we were able to ask conditional questions to provide a “testing board” for ranking policy options - “What proportion of pupils enrolling in elementary schools in 2027 will be enrolled with a deferral if only kindergartens could apply for a deferral on their behalf?”

For clarity, both types of questions will be referred to collectively as “unresolvable” questions below.

Of course, asking unresolvable questions could only contribute to our objectives if we were able to also obtain answers to them. A cursory glance at any of the large forecasting websites will show that conditionality and long resolution frames dampen user engagement, and so we had to devise a way to maintain a minimum level of activity even on questions which otherwise might be overlooked by forecasters.

We thus modified our reward structure so that final placement rewards were contingent on providing forecasts for at least 75 % of questions posed throughout the tournament. As our initial expectation was that unresolvable questions would make between 25 and 40% of all questions (the final proportion was 35 %), this threshold was sufficiently lenient while still requiring the answering of at least some unresolvable questions by all vying for the final rewards.

While this modification by itself could be expected to raise the engagement on unresolvable questions, it would do nothing for the expected quality of responses. Therefore, we added another layer: after the end of the tournament, a quarter of the unresolvable questions would be randomly selected and submitted for consideration by a group of subject experts who would indicate their own probabilistic expectations for the question. The average of those would then be used as the resolution datapoint. The experts would have at their disposal all rationales written by forecasters on that question. With this, forecasters were incentivized to provide thoughtful responses, so that they would not risk dropping out of the top ranks.

Unfortunately, the random allocation of unresolvable/resolvable questions in time[8] does not lend itself to easily interpretable relationships between the inclusion of such questions and participation. However, we are able to conclude that at worst, we did not see significant differences between the least-engaging resolvable and most-engaging unresolvable questions.[9]

As we considered ways to limit tournament participation drop-off, one of our interventions was reducing the number of unresolvable questions, and making sure that no challenge was composed purely of such questions. Overall, however, user feedback to unresolvable questions took a form slightly different to what we expected. While our expectation was that the main concern voiced would have been the inherent unpredictability of such questions, comments primarily focused on the mechanism for which a scored "baseline" would be obtained.

Three-week question lifetimes

Another aspect in which our tournament was rather unorthodox was the use of short time-frames for the provision of forecasts. As all our questions were parts of three-week challenges, there were some occasions where forecasters were asked to consider outcomes for which the ground truth would be known within the timeframe of the tournament, but outside the timeframe of the challenge itself.

Of course, this wasn't the case because of our disbelief in the effectiveness of updating predictions. Rather, this is the way we chose to address one of the fundamental difficulties of embedding forecasting in the policy process. This is the fact that ultimately decisions need to be made at a specific point in time, which means both that:

A. The aggregation of as many viewpoints as possible by the cutoff date is of higher value than subsequent updates, and
B. The marginal utility of updated information after a decision has already been made drops rapidly.

[8] There were two weeks with resolvable questions only, and one week early in the tournament with just unresolvable questions.
[9] Which is not to say that unresolvable questions were always and entirely less engaging than the other.

Both of these facts led us to adopt a design whereby we incentivize the output (aggregated forecasts) to be as up-to-date as possible and contains as many points of view as possible, while making sure that the quality of the most recent forecasts is not jeopardised by more effort being spent on the updating or refinement of forecasts which have already been used in the decision-making process.

Some members of our team are active forecasters themselves. Therefore, we were sympathetic to the argument that being scored when a majority of the time between the initial forecast and the resolution being known is closed to updating may not be the best user experience. However, there are three main reasons why we nevertheless settled on the described approach:

1. By working with (predictable) timeframes and clear deadlines, we maximised the number of involved forecasters while still reflecting the urgency of the information therein;
2. It affected every forecaster equally, so that rewards (which in all aspects were based on relative, rather than absolute, performance) were not impacted in this way;
3. With the closing and resolution dates known, the possibility of further developments in the time between them could already be reflected in the forecast itself, e.g. by widening the confidence interval appropriately.

While far from a perfect solution, then, we feel reasonably confident that this approach was the best fit given the circumstances, resources, and scope of the project. More broadly, our experience with this shows that structural interventions into the policy process itself may be critical for sufficient take-up into public administration. After all, feeding a (chronologically) continuous input into a (chronologically) discrete process is a challenge which goes beyond the scope of familiarising policymakers with probabilistic information.

Participants

Recruitment

We recruited participants for the tournament in several ways. Firstly, we utilised the contact list from our previous forecasting tournaments OPTIONS, to invite Czech residents who have already shown interest in forecasting earlier.

Secondly, we ran a small, targeted ad campaign on social media, especially on Facebook and Linkedin, where our audience is the strongest. We also renewed our old OPTIONS Twitter account, renamed it accordingly and used it to spread the information about our new initiative, as the followers were already interested in forecasting. This complemented a wider publicity push, which included even an interview with a call to action in a popular online magazine.

Finally, there were two demographic groups which we reached out to individually - women and college students.

While women were among the best-performing forecasters in our previous tournament, their overall representation both there and in the lead-up to FORPOL was lower than we found adequate. This was an especially strong concern in FORPOL, where the outputs were to be used by policymakers, making the questions of equity and giving a voice all the more important. To make sure the tournament's call to action message was sufficiently circulated among potentially interested women, we partnered with a network of women scientists to share information about the tournament to their members and on their public social media. Members of the women scientist network also helped formulate our claims better, to make them more inclusive.

For students, we conducted an outreach by a comprehensive flyer campaign distributed across multiple university faculties and various social media platforms used by university students. The primary objective was to inspire students to enhance their skill sets, potentially yielding substantial benefits in their personal lives, academic pursuits, and prospective professional trajectories.

OPT

One further way of establishing continuity with regards to our previous tournament was the inclusion of the OPT - expert prediction team. We have reached out to its members specifically as we were keen on providing other (non-OPT) forecasters with an additional motivation emanating from the fact that they have an opportunity to measure themselves against the best individuals from our previous tournament. Out of 60 OPT members, around 25 have chosen to participate in the tournament. As a group, they achieved a better aggregated score than those new to forecasting, further cementing their ability to outperform the crowd.

At the end of the tournament we were able to also identify a new batch of participants demonstrating advanced forecasting skill. First 30 of them were then invited into the OPT, and 26 of those decided to join. Their motivations for joining were diverse, but by far the most prevalent one was their self-development. The rationale behind the continuation of the OPT concept even after the tournament's end is as follows: The ability to identify forecasting talent has been one of the main focuses of our project, as we have encountered some distrust during our local outreach activities regarding the inexperience of participants and their perceived unreliability. With a proven track record and prestige, the OPT can now serve as an instrument well-suited to answer such types of questions more reliably and be better perceived by partners, who are now aware that it is a group of motivated, educated, skilled and experienced forecasters keen on understanding the world and predicting its future developments.

Mental/cognitive aspects

Motivation

At the outset of the tournament, we expected there to be three main sources of motivation for forecasters participating in the tournament. Namely, we thought of participation being driven by altruistic (helping policymakers), competitive/financial (outperforming other forecasters) and/or self-development (skill acquisition) motivation.

Looking back, we overestimated the attractiveness of financial incentives and underestimated another aspect.[10] While we initially thought notions of self-development might fall under the general purview of competitive motivation, we found that this is not the case. A substantial number of participants saw them as quite separate altogether, and it was the key to mobilising the most active participants (see final section of this chapter).

Drop-off

The tournament experienced significant drop-off in participation rates over time. Across the full six-month duration, we dropped from under 200 to 30-40 participants towards the end. Interestingly, very similar rates were reported[11] in the Salem Center/CSPI Forecasting tournament, which overlapped in time with FORPOL.

While this suggests that these dynamics are within the norms of comparable tournaments, we still believe it is productive to discuss three aspects of the tournament which may have accelerated the drop-off, especially in the context of their interplay with the above identified sources of motivation.

Length of tournament

Firstly, the length of the tournament itself might have dissuaded some participants. However, as detailed earlier, our initial recruitment was already skewed in such a way that most of the participants were interested in forecasting, so the effort itself was likely not the main cause of drop-off in these scenarios. Instead, what we gathered was that a significant barrier was the **sustainability** of these efforts. In other words, tournament length - of which all participants were informed at the start - only became an issue once it came to be used as a proxy for the effort necessarily expended before rewards (irrespective of their type) would be distributed.

[10] Relatedly, Metaculus has observed that prize pools alone offer only modest motivation. Forecasters are also motivated by fun and interesting questions, self-improvement in thinking rigorously about the future, competition, and the opportunity to inform decision making. Still, Metaculus notes that counterfactuals are hard to assess and they expect that prize pools, especially larger ones, are likely to help compensate for the effects that highly technical questions can have on engagement.
[11] https://www.cspicenter.com/p/new-5000-prize-in-the-salem-centercspi

Complexity of questions

The effort forecasters imagined necessary for the tournament was then clearly magnified by the second critical aspect - the complexity of questions posed. As mentioned earlier, throughout the process of developing questions with each individual partner, we were mindful of the long-standing difficulties in navigating the rigour/relevance frontier.[12] However, in following this avenue, we occasionally stumbled onto questions which were strong on both accounts, yet in doing so became all but indecipherable to forecasters without significant amounts of background knowledge.

Unfortunately, even if we endeavoured to provide or signpost such knowledge in the question descriptions, we suspect a strong chilling effect of these questions on participation, even though individually within their respective "challenges," there were only a few significant outliers.[13] While the "selfish" motivation factors may not have been impacted by this as much (as a minimum number of questions to be answered to be eligible for rewards), the responsibility felt by those who primarily viewed their participation in altruistic terms could have dissuaded them from participating on such questions. "Better not to participate at all," such thinking would go, "than to bias their good responses." While forecasting tournaments can hardly claim immunity to the "garbage in-garbage out" adage, what is striking is the seeming overlooking of the benefits of forecast aggregation.

So while this experience does speak to the need to pay more attention to the accessibility of questions, it also points to another aspect for further improvements. Greater attention must be paid in subsequent projects to clearly elaborating the assumptions behind why participation will be helpful.

Long-term questions

Thirdly, "selfish" motivations (financial rewards, competitiveness, and self-development) may have instead been dampened by the inclusion of long-term or conditional (unresolvable) questions. While in the case of the former two, the effort payoff is (seemingly arbitrarily) subjected to further variance, in the case of self-development, we assume that such questions can be seen as breaking meaningful feedback to forecasters' cognition. In either case, they distort the expected value of participation on such questions.

As we monitored participation dynamics throughout the tournament, we tried modifying some parameters (i.e. reducing the complexity/frequency of questions, proportion of long-term

[12] Discussed not only in the academic sphere by i.e. Philip Tetlock, but also noted for example in a similar context to ours by Ben Roesch of Cultivate Labs: *"In addition, most forecasting tournaments involve an inherent tension between the questions that decision makers want to ask (often hard or esoteric) and the questions that interest forecasters. Any forecasting effort should dedicate significant attention to the question portfolio, making sure to strike a balance between "hard" questions and engaging ones."*

[13] Interestingly, the most obvious outliers when controlling for time relate to the education questions in challenges 5 and 9.

questions) to increase user retention. However, these did not seem to significantly alter the dynamics. Based on the above, we would consider two more structural interventions to pair with any future projects in forecasting:

- Providing more detailed guidance on forecasting effectively. While we provided a short crash-course on calibration and the principles of forecasting (i.e. base rates), knowing what we do now, we would focus more on a well-designed approach to cultivating the self-improvement motivation. The period between forecasting and receiving feedback can already be quite long as is, so intermediate steps to enable forecasters to feel like they are improving can be of great value.
- Improving the discoverability of information on forecasting-bundled concepts, such as crowd wisdom, Bayesian updating, etc. While partners were surprisingly little interested in peering into the forecasting black box, forecasters themselves might be mobilised better if they can more readily discern a justification for their participation regardless of the particular situation.

Evaluation interviews & Feedback

Forecast enthusiasts

In March and April, we organised a series of interviews with the best forecasters from the tournament (OPT members and new forecasters with solid scores alike). Through these, we have not only learned how significant the self-development motivation is for them (way above our expectations), but they also mentioned that the financial incentive only creates a more prestigious perception of the competition, but in the end is not a decisive factor for them. Generally, they were not focused on a single "favourite" topic, but rather preferred to be challenged intellectually in various ways: they welcomed diverse themes as this meant they had to research and prepare before giving proper predictions and nudging them to expand their knowledge base on current issues and dive deeper into their existing assumptions.

Importantly, none of the participants interviewed ranked improving the preparedness and policy planning of public institutions as a crucial factor for their participation. They agreed on seeing it as a "nice bonus," not something that would captivate their attention and single-handedly push them towards active participation in the tournament.

Drop-outs

In April 2023, we held several interviews with forecasters who became inactive after several months of participation in the tournament, despite being on track for top ranking spots until then. We were interested in hearing their reasons for dropping out, so that we could design our programmes (one of the goals of which is to identify good forecasters, after all) to keep such participants engaged.

All interviewees noted that they appreciated the number of topics covered by the tournament and highlighted the lack of time - of which significant investment into the tournament was necessary, in their view - as the main cause for ceasing participation. One of the participants

decided to re-engage in the tournament after the feedback interview and later was invited into the OPT based on his results.

Polls

Based on the results of the final poll at the end of the tournament active phase, some of our other findings were further reinforced. Of 15 participants, 14 selected self-development among primary motivators, with fun/entertainment coming in second, with 9 responses.[14]

All the participants filling out our feedback questionnaire said they would recommend forecasting to their friends and colleagues (the lowest score was a 7 on a 0-10 scale). 12/15 participants said they plan on being active forecasters after the tournament, with the remaining 3 left unsure at time of polling, primarily due to time constraints.

For most participants (11/15 who filled the questionnaire) it was interesting or motivating that the results were directly communicated to our partners from civil administration institutions, however, it was not among their main motivational drivers (only 4/15 selected this option). Forecasters also provided us with a number of concrete recommendations to further improve our forecasting tournaments, most of which included organisational aspects such as more detailed upfront elaboration of the scoring mechanisms.

Forecast Data

Although the data we have collected on questions and forecasts throughout the tournament by no means represent a massive dataset, we nevertheless present below some findings and tentative conclusions of note derived from a quantitative analysis of data provided by Metaculus.

On approaching the data

Before we share some of our findings, a few sentences are in order regarding the used concepts and available data. Over the course of the tournament, 8 027 forecasts were submitted to the Metaculus subdomain. For the bulk of the analyses presented below, we started our inquiry with this data, structured along the following columns:

- Question identifiers - the ID, group question ID if applicable, long-form question title
- User identifiers - the user ID and username of the forecast author
- Question type and range - the type of the question (continuous/binary) along with information regarding its answer scale: whether the upper and/or lower boundaries

[14] Though our sample was even more self-selected than theirs, these results align with a community survey of Metaculus users from Q1 2023, where the top four reasons for visiting Metaculus were “Train myself to think rigorously about the future”, “Have fun”, “Learn about fascinating topics”, and “Keep up with the news.” See https://www.metaculus.com/questions/16626/-sharing-metaculus-community-survey-results/

were open or closed, what their respective values were, and whether answers were submitted on a log scale (not relevant during FORPOL)
- Prediction content - a representation of the probability density functions (and possible factors) submitted by users using the Metaculus UI
- Prediction time - a timestamp of the forecast's submission

Based on these, we also created a further variable used across the analyses presented below:
- Activity sessions - for every user, we sort their predictions by time, and then group them into sessions. Sessions start after the end of the previous one or with the first prediction, and then run for 90 minutes (a rough estimate of our team on what amount of activity could still be considered part of the same block). As such, sessions give a rough estimate of user activity in FORPOL without being as sensitive to outliers as i.e. the number of forecasts submitted.

User engagement

The number of engaged users dropped steadily over the tournament's timeframe, as noted previously. However, towards the end, user engagement seemed to reach an asymptote and hovered around the same values.

Drops in engagement came primarily from drop-off in between challenge rounds, with the most notable drop being between the first and second challenges, where there is no overlap between the range of users engaged in the former and the latter.

Predictors - question-side

Based on theoretical assumptions and the availability of data, we constructed further variables for each unique continuous[15] question in the tournament:
- days since the start of the tournament (*days_since_start*),
- whether the question was unresolvable (*longterm*),
- whether the answer scale used in the question had zero, one or two open boundaries (tails) (*tail_count*), and
- the amount of freedom when answering (operationalised as the ratio of the available answer space to the value of the upper bound[16]) (*scale_prop*)

These variables represented characteristics of the questions which we identified as possible factors in driving or depressing engagement with questions

[15] Many of the considered predictors could not be harmonised to work for both binary and continuous questions at the same time. Therefore, we decided to conduct a first search for statistically significant predictors on continuous questions, which are amenable to a greater range and also provide a larger dataset. If we later found that the predictors that turned out to be statistically significant could all be transferred without issue to binary questions as well, we would repeat this process with all questions and shortlisted predictors.

[16] Even with open tails, questions on the Metaculus platform have an answer space with a lower/upper bound. Resolutions beyond these are treated as a binary question on the crossing of that bound.

Afterwards, a linear model was constructed using the *lm* function of the *stats* package in R. The following table shows the coefficients and p-values assigned to each variable tested for this model:

Table 1: Coefficients and significance of potential predictors of question engagement

Variable	Coefficient	p-value
days_since_start	-0.6199286	2.682588×10^{-15}
Intercept	*143.4553743*	2.973695×10^{-14}
tail_count	-11.9415766	1.369777×10^{-2}
longterm	-9.6373976	6.239606×10^{-2}
scale_prop	-12.2712201	2.756284×10^{-1}

Two of these predictors show with a statistical significance under p = 0.05. The strongest evidence is found, unsurprisingly, for the impact of the number of days elapsed since the start of the tournament.

Secondly, quite surprising was the rather strong statistical significance and coefficient of the *tail_count* predictor, tracking the number of open boundaries in the particular question. In other words, it would seem that adding an open boundary to a question significantly impacted its engagement levels. Our data don't allow us to draw any further conclusions based on this, but some preliminary ideas from our team include that closed boundaries could have contributed to forecasters feeling more "in control" of the outcomes of their predictions, as they could model their submitted probability distributions across all (or more than would be the case with open boundaries) possible resolutions.

Barely above p = 0.05 was the coding for unresolvable questions - and in any case, the coefficient being the second smallest (behind, understandably, the number of days since the start) did come as a surprising result to us, as our impression based on user interviews was that this could have been a substantial factor.

Finally, the *scale_prop* predictor does not seem (at least linearly) tied to engagement - not only does it exhibit a very high p-value, including it raises the model's AIC[17] slightly (from 248.46 to 249.11). This indicates that considering this variable in the model actually lowers its predictive capability.

[17] The Aikaike information criterion (AIC) is a widely used estimator of the predictive quality of statistical models, which can address both the risks of under- and over-fitting. The preferred model is that which leads to the lowest AIC value.

Predictors - user-side

At the same time, some of our user recruitment channels allowed us to solicit data on users' demographic profiles - age, educational attainment, and region of residence. As a result, more than half of the submitted forecasts can thus be matched with all of these. Additionally, for some users we also obtained information on their academic area of interest, although this was not always applicable and is thus available less frequently as a data point, so we could not at this stage operationalize it for analysis.

Whilst we tested several further models, none of the above characteristics appear to demonstrate any significant effect on the number of questions answered or the total number of predictions submitted by a given user. If anything, once we already select for users which submitted this data, linearly modelling the effect of age also increases the AIC.

Updating

Unsurprisingly, our forecasters tended to update (i.e. submit a forecast on a question where they already submitted at least one other forecast before) less rather than more. In total, 26% of unique user/question combinations saw updates, with almost two thirds of forecasters updating at least once during their participation in FORPOL.

In the full forecaster population, fifteen users updated on more than 90 % of the questions they participated in. At the same time, however, some two fifths avoided updating altogether (36,6 %) or updated only extremely rarely (approx. 3 % with updates on less than 8 % of their questions).

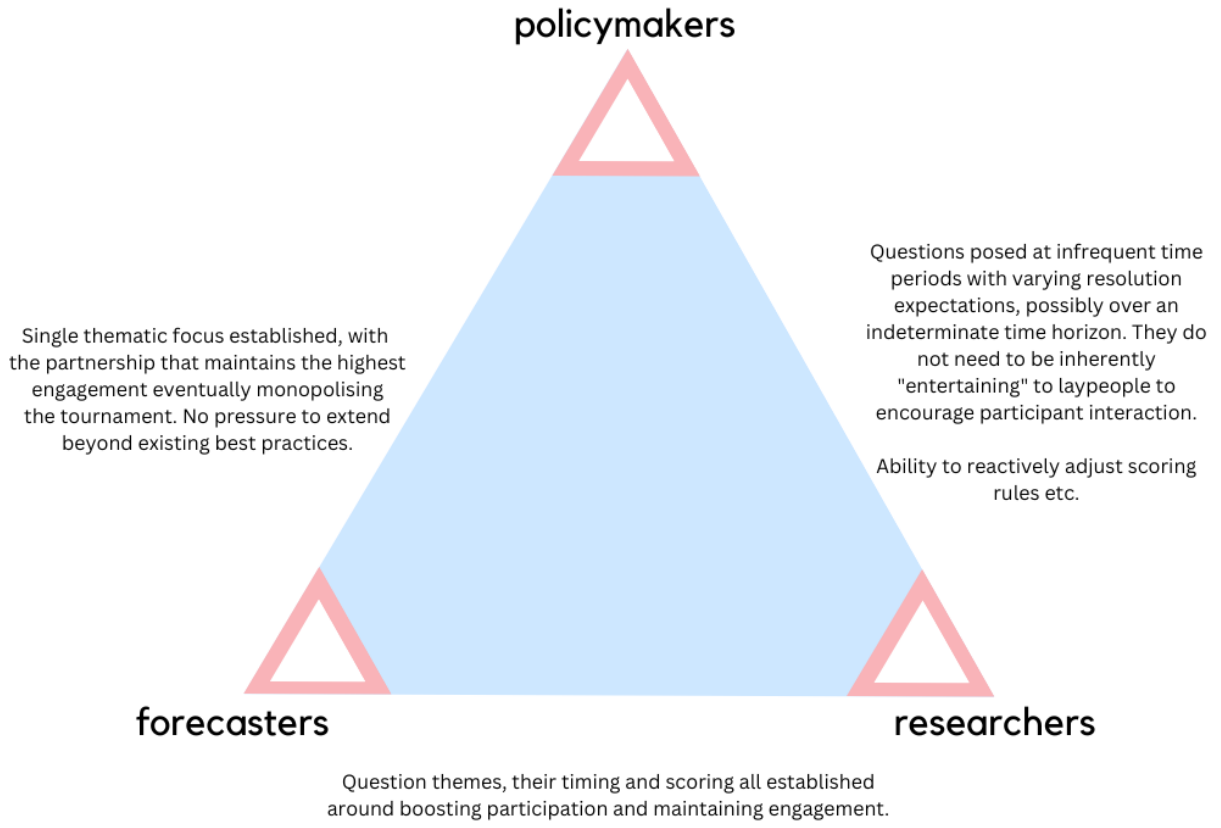
When filtering to only those users which recorded at least two different sessions in the tournament, however, although the number of extremely update-prone users drops slightly, we notice a more significant decrease in the number (and proportion - 28 %, of which 24,5 % with no updates).

Notably, there was no significant relationship found between the popularity of a question and the propensity for users to update (i.e. there was no "competitive" drive to update more frequently) - the p-value for a Pearson correlation test is 0.86, with a sample estimate of the coefficient of 0.025.

Reflection

Throughout the project, one particular issue that kept resurfacing in distinct ways was the trilemma we faced between our responsibilities to:

- forecasters, for whom FORPOL was an entertaining competition, but one which had to remain transparent and predictable, as well as manageable with limited time commitments - without whom no forecasts would emerge;
- our partners in public institutions, who sought answers to vexing questions which may struggle to be used for forecasting - without whom forecasts would not be policy-relevant; and
- forecasting researchers, which had always to be kept in mind with regards to the interaction of the former two stakeholder groups, yet sometimes gave rise to wholly new requirements or challenges, such as during partner selection or question generation. Failing to pay sufficient attention and foresight to this factor could then endanger the entire research project.



From this experience, we take that any further research should seek to minimise the tension between these, ideally by redefining the relationship to one of these stakeholder groups. At this stage, the most prospective way would be to change the nature of the “commitment” towards the forecasters. If a group may be found that is more dedicated and motivated to provide forecasts beyond “fun” and a prospect of ranking well enough to claim prize money, the trilemma could be reduced to a much more manageable, less frequent dilemma.

We can also identify with the experience presented by Cultivate Labs, who write *"We've been running crowdsourced forecasting programs for years and while we've had our share of successes and failures, none of them have failed due to inaccurate forecasts or issues with scoring."* In any case, even some of the difficulties identified above relating to accuracy or scoring issues were the product more of specific types of communication or expectations by forecasting laypersons rather than fundamental issues with the mechanisms themselves. Equally, the feedback we received from our partners did not point towards them questioning the effectiveness of the method or the track record of forecasters being a significant obstacle.

When it comes to engaging with the selected members of the OPT team and its partners, after the tournament's conclusion, we have formalised the OPT community. As we learned that it is important to have both clients and forecasters on board before the questions are asked, we held brainstorming sessions to identify topics that resonated well with OPT members.

Moreover, we are now actively working with our forecasters to leverage their own contacts and share their ideas to give them the opportunity to influence the direction the OPT is headed. We are very optimistic about the potential of the OPT, as initial talks have shown that both past and potential policy partners are even more interested in the concept - with every single one of our partners providing feedback asked to be contacted and kept in the loop with further developments.

Throughout the project, we noticed several generalizable aspects of both working with policymakers and designing forecasting tournaments. The most notable of these are summarised in the executive summary of this write-up as lessons learned. In general, we can say that our experience shows that the potential demand for forecasts at large can be significant. This means that it is possible to pay increased attention to the types of collaborations embarked upon - with selection for the most prospective endeavours, as forecaster attention and engagement is far from an unlimited resource. Furthermore, the types of monetary and other incentivisation that work best for these goals remain an open question, and should be subject to further investigation.

References

Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., ... & Mellers, B. (2017). Distilling the wisdom of crowds: Prediction markets vs. prediction polls. Management science, 63(3), 691-706.

Howlett, M., Giest, S. (2015). Policy Cycle. In: James D. Wright (editor-in-chief), International Encyclopedia of the Social & Behavioral Sciences, 2nd edition, Vol 18. Oxford: Elsevier. pp. 288–292.

Roesch, B. (2023). Does forecasting accuracy really matter? Cultivate Labs Blog. https://www.cultivatelabs.com/posts/does-forecasting-accuracy-really-matter [Accessed on 29 May 2023]

Samotin, L. R., Friedman, J. A., & Horowitz, M. C. (2022). Obstacles to harnessing analytic innovations in foreign policy analysis: a case study of crowdsourcing in the US intelligence community. Intelligence and National Security, 1-18.

Satopää, V. A., Salikhov, M., Tetlock, P. E., & Mellers, B. (2021). Bias, information, noise: The BIN model of forecasting. Management Science, 67(12), 7599-7618.

Savelli, S., & Joslyn, S. (2013). The advantages of predictive interval forecasts for non-expert users and the impact of visualizations. Applied Cognitive Psychology, 27(4), 527-541.

Sempere, N., & Lawsen, A. (2021). Alignment problems with current forecasting platforms. arXiv preprint arXiv:2106.11248.

Stastny, B. J., & Lehner, P. E. (2018). Comparative evaluation of the forecast accuracy of analysis reports and a prediction market. Judgment and decision-making, 13(2), 202-211.

Tetlock, P. E., Mellers, B. A., & Scoblic, J. P. (2017). Bringing probability judgments into policy debates via forecasting tournaments. Science, 355(6324), 481-483.

Annex I:
Questions posed in the tournament

1

What will be the Czech Republic's reported gas storage levels on 9 January 2023?
Will a Prime Minister of the Czech Republic be received in Washington, D.C. on an official visit to the U.S. by 31 December 2022?
Which candidate will the ANO party support in the first round of the 2023 presidential election?
How many members will the SPD party's parliamentary caucus have as of 30 March 2023?

2

How many persons will the Czech Ministry of the Interior grant temporary protection status to due to the war in Ukraine by 31 December 2022?
In which week between November 2022 and January 2023 will the largest increase in persons granted temporary protection status due to the war in Ukraine occur?
What will be the ratio of mortgage loans originated in Q4 2022 for acquisition by construction compared to mortgages for acquisition by purchase?
According to the next PAQ Research report, what proportion of Ukrainian refugees previously accommodated in hostels will be renting or living in separate apartments provided by Ukrainians or Czechs?

3

How many Czech cities with populations between 20,000 and 100,000 on 1 January 2022 will have more than 10% fewer residents on 1 January 2032 than on 1 January 2022?
By 2050, will the Czech Republic experience floods with estimated total damages in excess of 120 billion in January 2022 prices?
How many inhabitants will the municipalities of the Prague metropolitan area have on 31 December 2030?
In how many years between 2041 and 2050 will the number of cooling degree days exceed 100 in at least 10 regions of the Czech Republic?

4

When will Andrej Babiš arrive at the 17 November monument on Národní třída in Prague on 17 November 2022?
How many people will take part in the demonstration organised by the PES Movement planned for 17 November 2022 on Letná Plain?

5

What proportion of pupils enrolling in elementary schools in 2027 will be enrolled with a deferral?
What proportion of pupils enrolling in elementary schools in 2027 will be enrolled with a deferral if only kindergartens could apply for a deferral on their behalf?
What proportion of pupils enrolling in elementary schools in 2027 will be enrolled with a deferral if the number of children per teacher in kindergartens was reduced by one quarter by the start of the 2027/2028 school year?

6

What percentage of Czechs with an established eID will have used it at least once by 12/31/2029?
What percentage of citizens in rural areas of Czechia completed an online education course in 2022?
What percentage of Czech residents over 16 years of age will be living in households without internet access in 2029?
Will ATM withdrawals account for more than 12% of the total number of credit card transactions in the country in 2022?

7

How many PCR test results for Covid-19 disease will be reported in the Czech Republic in the first three months of 2023?
How many cumulative man-days will be spent by Covid-19 patients in ICUs in the Czech Republic in the first three months of 2023?
How many doses of approved Covid-19 vaccines will be used in the Czech Republic in 2023?

8

How many Ukrainian citizens will the Ministry of Labour register on the labour market in February 2023?
What will be the ratio of job applicants of Ukrainian citizenship to employed workers of Ukrainian citizenship in the Czech Republic in February 2023?
By 31 March 2023, will Škoda Auto or the Volkswagen Group announce that the plan to build a so-called Gigafactory in the Czech Republic will be implemented?
By 31 May 2023, will the Czech Prime Minister or any four members of the Czech Government resign their positions?

9

What will be the total volume of applications for support of school special education teachers and psychologists by primary schools and pedagogical-psychological counselling centres under OP JAK by 28 April 2023?
How many vacancies for school principal positions will be reported by the pedagogical community with a submission deadline between 1 January 2023 and 31 May 2023?

How many kindergarten and elementary school teachers who taught in 2019 were no longer teaching as of 1 February 2023?

10

What percentage of Czechs will report they find it "difficult" or "very difficult" to make ends meet with their household income in the spring of 2023?
What percentage of Czechs will say they believe that the security of the Czech Republic is "definitely" or "rather" threatened in the spring of 2023 under the prevailing international situation?
What percentage of Czechs will answer that most people "rather cannot" or "definitely cannot" be trusted in the spring of 2023?

11

What percentage of companies exhibiting at the COFIT Career Fair in March 2023 will dedicate some of their materials to women in IT?
What will be the percentage of attendance at Czechitas face-to-face IT courses to total attendance at all courses in January 2023?
How many accreditations will be granted to advanced IT retraining courses in the Czech Republic in 2027?

12

Will the annual inflation rate in the Eurozone fall below 5% by the end of May 2023?
Will the German 10-year bond yield (TMBMKDE-10Y) fall below 1% by the end of May 2023?
What will be the closing price of the 2024 Slovak electricity contract (Cal-24) in the Baseload category in EUR/MWh on the EEX-PXE on 31 August 2023?

13

According to the TradingEconomics forecasting model, what will be the value of the LMEX index in 12 months' time on 31 March 2023?
What will be the percentage change in OECD countries' foreign trade volume with China in inflation-adjusted USD in 2029 compared to 2019?

14

How many persons will file for personal bankruptcy in the Czech Republic from 1 January to 31 May 2023?
How many distraint warrants will be granted in the Czech Republic in 2023?

15

What percentage of the population of the Czech Republic aged 25-64 years participated in some form of lifelong learning in 2022?
What proportion of Czechs who are interested in participating in lifelong learning activities in 2023 will identify lack of finance as one of the main barriers to participation?

16

What will be the number of project proposals submitted in the final Call for Proposals of the TREND R&I programme?
What will be the success rate of domestic actors in the 9th EU Framework Programme for Research and Innovation (Horizon Europe) during its duration, i.e. 2021-2027?

17

How many forensic experts will apply for registration in the new register by 31 May 2023 in order to continue their activities from 2026?
How many forensic experts will apply for registration in the new register by 31 May 2025 in order to continue their activities from 2026?

18

How many housing allowances will the Ministry of Labour disburse in the first four months of 2023?
How many illegal crossings of the EU's external borders will Frontex record in the first four months of 2023?

Annex II: Methodological Guidelines

Implementing a forecasting tournament for public policymaking comes with specific challenges, which are best considered in advance. We created this simple methodological manual[18] to explain and simplify the whole process.

Before any suggestions are presented, however, a note on how to use this document. Firstly, we can only offer detailed comments on the task which we ourselves engaged in - running tournaments to identify accurate forecasters and build trust and interest among policymakers. We did not have explicit goals of establishing long-term cooperation with the institutions, creating early warning systems, etc. As a result, we cannot offer tried and tested advice on these efforts. This should not be taken to say, however, that we somehow view them as having less importance or potential value.

The approaches and goals chosen for the FORPOL project responded to the situation and research questions we wanted to address. Some steps (such as building a community and identifying the best-performing forecasters) are a necessary precondition for other projects (such as building teams of forecasters). If you are considering a substantially different project, we however still believe that the contents of the following sections phrased as questions can help guide some aspects of your design. **While we cannot speak as to what the best choices may be, we can share some questions which can point to a good way forward.**

When planning a project to develop crowdsourced forecasts directly for the use by policy- and decision-makers, we find it helpful to think in terms of the following stages:

1) Choosing the platform:

A forecasting platform is of critical importance for the success of any crowdsourced predictions. It builds confidence in impartial score-keeping and feedback, but also acts as the environment which is to be engaging (and user-friendly!) for forecasters. As noted above, we knew that a tournament structure (as opposed to a prediction market) would be more appropriate to our objectives very early on, and we expect this to be the case for most similar projects. Three key reasons for this are:

- Already preliminary discussions with policymakers in advance of the project pointed out that some interesting and important policy questions would struggle to be expressed in market terms. Furthermore, in extreme cases their use might also create unnecessary reputational risks when interpreted by laypeople as “betting” on bad outcomes.
- Furthermore most policy questions for which external crowdsourcing may be appropriate (at least in our experience) - i.e. not operation-critical or concerning

[18] This section shares a bibliography with the main write-up text.

issues such as immediate national security - tend to relate to longer-term time horizons (several months to several years). For such questions especially, prediction markets appear to have the biggest relative advantage over prediction markets (Atanasov et al., 2017)

- Finally, for the policy-forecasting nexus, questions about the accuracy, efficacy, or other aspects of tournaments can be much more comprehensively answered - the literature is quite broad, with many research programmes and fields of inquiry advancing in parallel.

There are numerous forecasting platforms to choose from (with the most well-known providers being Metaculus and Cultivate Labs), each of these with specific pros and cons. There is, however, a rather standard suite of functionalities for such platforms - submitting predictions, discussion boards, viewing the evolution of the consensus prediction in time, etc. Some of the differences include i.e. the way results are presented to forecasters (for their feedback) - is the focus on proprietary points, or i.e. a standardised metric such as the Brier score? Similarly, some platforms are oriented towards forecasting through distributions, while others operate on the basis of bins - but we have not registered a clear preference among policymakers for either of these methods.

Some projects may be initially designed with additional aspects in mind - ones that no existing platform can support at the moment. While it may seem tempting in such cases to develop basic platforms to support such aspects, our experience would urge caution, and in any case point out one critical issue. Establishing scoring rules which properly align the incentives of forecasters to the ideal goals of the forecast exercise is far from a trivial issue, as was comprehensively demonstrated by Sempere and Lawsen (2021). Established platforms, in the meantime, have developed know-how on building on top of proper scoring rules to develop further incentives. Building such systems from scratch may present significant inefficiencies in many cases.

2) Cooperation with institutions (planning):

This stage should be used to plan detailed cooperation with public institutions. There are plenty of questions to answer and many decisions to make e.g.:

- What are your main goals? Our own project ambitiously strove to advance three different goals: improving public understanding and acceptance of forecasting; identifying and developing top forecasters; and crowdsourcing forecasts useful for public policy. What we did not appreciate sufficiently at the outset of our project was the frequency with which (some of) these would clash. Try to be clear up-front about which of these are critical for you, and think proactively about how you will prioritise them should the need arise (i.e. “Will we ask additional questions partners are keen on, even at the risk at forcing drop-off in users when they see a large number of questions posted at the same time?”)

- How long will the cooperation(s) be? In our experience, it usually took between 8 to 10 weeks from an initial meeting to the delivery of a report with predictions - although this assumes questions can be posed shortly after they are developed. Moreover, it is very likely that even after those 10 weeks, the true resolution date for many of the questions will be far in the future. Long-term cooperation (such as repeat question batteries developed with a single partner) might improve buy-in by creating additional opportunities to discuss the results and use of previous predictions. However, **one of the most critical limiting factors is forecaster attention/fatigue**. Thus the number of questions (and, subsequently, the "depth" and "length" of partnerships) must be consciously managed. It is also worth thinking in advance about the scheduling of questions to keep both partners and forecasters engaged.
- Do you want to focus on a specific field of public policy? Narrowing the thematic focus of a forecasting tournament may increase the speed with which you are able to establish a "foothold" in the relevant policy domains. However, it is worth thinking about the risk that the method then becomes associated with those domains, at the expense of others. Similarly, in our experience, forecasters were drawn by the diversity of questions, and have strongly indicated their participation would have been lower in its absence. If you have strong reasons for topic-specific tournaments, we suggest at least cooperating with different leads/departments/organisations within, to establish a different type of diversity.
- How many partners should the project have? A smaller number of partners lowers the administrative burden of repeated introductory cycles and the associated costs. However, a lack of sufficiently diverse question topics may have a negative effect on participation. Moreover, limiting the tournament to a few departments may have detrimental effects on supporting desired analytical methods (updating, keeping score) across public institutions.
- How are you going to reach them? We were able to build in a large part on pre-existing relationships with ministries and government agencies. However, this may not be the case for every project team. To help, we developed the **Practical Explainer** published with these guidelines. This synthesises what we found worked best or what most often asked when initially meeting less familiar potential partners. One more thing to keep in mind is that, in our experience, one of the greatest potential risks was creating bad/unfounded expectations. Don't be afraid to correct potential partners when you see a misconception arising.

3) Recruiting and motivations:

The motivation of each forecaster could be different and we recommend using a combination of all of them to target the recruitment of forecasters towards the widest possible group. As an activity on its own, forecasting has strong self-selecting tendencies already as is, and thus ensuring as many people as possible learn of the project and have the opportunity to participate if it sounds like something they might

take interest in is critical. From our own experience, top forecasters range from bank analysts and traders to small-town doctors, and fresh graduates to experienced industry professionals.

In our own recruiting efforts, we operated with four key motivators (listed in descending order of importance as voted by our forecasters):

- self-improvement ("come learn a useful skill and get feedback!"),
- sense of community and entertainment ("engage with other interesting people!")
- competition ("come win prizes for your efforts!"), and
- altruism ("come improve policy-making!")

The above list is corroborated by similar polling by Metaculus referenced in the project write-up. While we feel this information may help with prioritising messaging, we would still recommend not ommitting even the items lower on the list. This is both to maximise the tournament's attractiveness, and also to provide the most accurate possible idea about the tournament in advance. Related to this is our hard-won experience with **putting sufficient emphasis on the expected time requirements for participants**. As is described in more detail in the write-up, establishing clear initial expectations can help mitigate later factors and support the sustainability of a tournament.

During our recruiting process, we focused on students, women, young professionals and previous attendees of our tournament. We used multiple tools to recruit, from targeted ads and newsletters, through an interview in an online magazine, to flyer campaigns in Prague universities. Already in our experience, we have found occasional concerns about representativeness of the forecasting crowd, and we expect these to increase over time and in other (i.e. national) contexts. Especially in "pilot" projects, where the value-added can be easily communicated to policymakers in terms of the wisdom of crowds effect,[19] the diversity of crowds (which can then be thought of as the scope in which the crowd is capable of converging towards an accurate answer) will be of interest to partners, and that's without getting into aspects of equity and representativeness.

However, these motivating factors are critical not only when thinking about forecaster recruitment. They must be plausibly and practically translated into incentives for the actions which forecasters are to be incentivised to take. Most notably, key actions in a forecasting tournament (and the questions to ask in relation to them) include:

Submitting forecasts

[19] Although this line of argument is not limited to this framework and is equally valid when using i.e. the proposed Bias - Information - Noise (BIN) model (Satopää et al., 2021).

It goes without question that this activity of forecasters is the most critical for the usefulness of a tournament to public sector partners. However, the interplay of policy/decision-making and the sustainability of a tournament lead to several important decisions to be made regarding the form of incentivising the submission of forecasts:

- Will forecasters be rewarded in proportion to the number of forecasts they made?
    - If not, will a minimum level of activity be required for reward eligibility? If so, how will this be defined? - The decision taken on this can lead to significant path dependence for the tournament.
      While setting a minimum answer threshold can be easy to monitor, falling short of this can lead to compounding drop-off in the middle stage of a tournament (see write-up for more details). Moreover, it makes later sign-ups virtually pointless.
      On the other hand, establishing complex schemes (i.e. rolling minimums) might be difficult to monitor for both forecasters and the organisers in the absence of any purpose-built software.
- Will incentives for submitting forecasts focus only on rewarding based on eventual revealed accuracy?
    - While this is seemingly a straightforward way to incentivize the ultimate goal, it relies on the scoring rules of the tournament to be properly adjusted.[20]
    - If rewards will not be based also on other factors, which will be chosen? For example, the diversity of predicted topics, or the timeliness of predictions? Note that these need to be in line with the design of partner cooperation (for example, if partners only view the results once as a discrete action on a predetermined date, should early predictors be rewarded - and if so, to what extent?)
        - In any case, the mechanisms for this must be proactively communicated to forecasters to avoid misunderstandings emerging i.e. regarding the treatment of the timing of predictions etc.
- How often should incentives (rewards) be tallied up? For example, will tournament prize money only be disbursed in the form of overall grand prizes, or over time - i.e. for each cooperation or time period/round separately?
    - While the former is of course more robust and most likely to reward based on underlying skill rather than luck, especially in longer tournaments it may simply take too long between action and reward to form an effective feedback loop. In any case, partial prizes for accuracy should likely not be based on any fewer than tens of resolved questions.

Commenting

[20] See discussion and reference at the end of section 1 of this manual

As stated above, most current forecasting platforms provide functionalities for commenting to support one's predictions and share rationales/sources. While usually voluntary, recently there have been various ways across platforms of "nudging" users towards providing comments more often. In our experience, this is definitely something that should be promoted and likely incentivised directly, especially as the partners we worked with were very keen on delving into the rationales of forecasts.

- While the simplest way would be to simply require mandatory comments with each prediction (which some platforms even support as a global option), we would advise against this approach. In our experience, the significant decrease in average rationale quality is definitely not matched by their increased number, as many forecasters will just default to low information-density comments.
- Instead, we have found some success with **incentivising them with a secondary track of rewards**, which unlike accuracy rewards can be robust even when awarded intermittently based on limited samples.
    - Notably, in our feedback form, top forecasters in our project were very appreciative of this, and some even called for this track to form a greater portion of the overall reward pool at the expense of accuracy/participation rewards.

Updating

Due to the way our tournament was set up (see beginning of guidelines and write-up) - with discrete submissions of the "state of play" on forecasted questions to a certain date (relatively) shortly after the questions were first posed, we did not consider additional incentives for updates, as both current research and experience show that the desirable kind of updating is already covered by targeting forecast accuracy.

- More specifically, our theory of change operated under the assumption that often the act of policymaking involves locking in a course of action at an "arbitrary" point in time.
    - This deadline can be institutionally mandated even though waiting for longer may have provided more information or reduced uncertainty. In light of this dynamic and our goals, the chosen approach was appropriate.
    - However, other models of cooperation (such as early warning systems) may be designed, where the relative importance of updating rises significantly, and its incentivisation will be critical. Therefore, this point must be critically evaluated on a case-by-case basis.

One of the limitations of the public tournament format is the relative strictness of being bound by the initial rules under which forecasters have signed up. While it is necessary to establish these as transparently (both in terms of procedures and expectations participants may develop) as possible, it may be worth it to think about possible **appropriate clauses for minor modifications in the tournament's conduct**, especially regarding any supplemental incentives developed in responses to the dilemmas above.

4) Cooperation with institutions

The FORPOL project write-up offers first-hand description of the process we chose for developing partnerships and collaborating with public-sector institutions. They may be reviewed by those interested in more detailed descriptions. In this section, we will primarily address the take-aways from our experience.

Furthermore, an additional document - the **Practical Explainer** - has been developed, the second half of which may be read in conjunction with this section. It breaks down the cooperation on developing questions with policy partners into the key steps which we identified over repeated cycles.

A) Reaching out

To work in close cooperation with public institutions, their participation must first be secured. In our experience, policymakers were quite receptive to the prospect of being shown further impartial analysis and judgement on the issues they are concerned with. The most common issue seemed to be a general lack of time/attention on their end to dedicate to the cooperation. While not all institutions we contacted were ones we already had pre-existing relationships with, we were quite selective about “cold call” contacts. While the final success rate need not be significantly different between “old” and ”new” contacts, the latter may be initially more apprehensive. In our experience, among the most persuasive arguments were the facts that:

- The tournament itself is a system designed around the concept of feedback, so accuracy of predictions is targeted in this way;
- The partners drive the process with question selection and utilisation;
- Resulting reports contain the rationales of forecasters and offer guidance on how to think about the probabilistic information.

In contrast, we did not find significant pushback i.e. on the potential value of expressing beliefs in probabilistic terms. We tried to synthesise what we found to be the most helpful aspects of initial communication (offering case studies, pointing to relevant academic literature, and giving a clear idea of the expectations) in the Practical Explainer published along with these Guidelines.

B) Question development

Developing forecasting questions for use in the policy process should certainly follow general practice of question writing, as summarised i.e. in the Metaculus Question writing guide - such as keeping resolution criteria unambiguous or choosing appropriate references to determine resolution.

However, several additional aspects come with the specific task of co-creation of questions with policy partners. Though it may seem trivial, the type of questions appropriate for forecasting (i.e. “When” or “How many” - rather than “How best” or even “Why”) that can leverage its strengths must be continuously reaffirmed, as in our experience there is a tendency to approach the underlying “central questions.” This must be kept in mind and reinforced repeatedly. Next, there is the question of the appropriate format in which questions will be developed. One neutral approach, which worked in most cases in our project, is described in the write-up - although other, varyingly similar formats are also available. In our view, the selection of which will be most appropriate will be impacted significantly by (at least) two aspects:

- Has the client identified in advance a specific decision or deliverable to which the forecasts should provide inputs?
- How many questions can be posed with the partner in the tournament?

The dynamics of question development will no doubt have to be tailored to the answers to these - while the approach described in our write-up worked often, it is not necessarily one-size-fits-all, and others (such as the “Bayesian question clustering” first proposed in *Superforecasting* and then operationalised in INFER tournaments) have their own strengths.

Finally, there are three important aspects of any question development which take on specific characteristics in the context of a policy focus:

- Question types: Experience with running and participating in forecasting tournaments clearly indicates binary (% of yes/no) questions tend towards higher engagement. However, they are quite uncommon in policy questions. It may be tempting to circumvent this by posing some numerical questions as binary ones by setting a baseline which will/will not be exceeded. In such cases, however, it is important to discuss in advance the selection of this baseline with the partners and the interpretations of either result to minimise its arbitrariness
- Resolution horizons: The partners we worked with had the strongest preference for questions with a 7-month to 2-year resolution period. Based on the length of your project and the timing of question development (i.e. rolling basis, as was our case, or upfront), fitting into this window may be difficult, and it may take some effort to find appropriate questions in this regard. Some ways to address this include considering using the platform for “nowcasting” (i.e. reducing uncertainty about current states which are unknown, crucial, and will become known sooner than future ones) or reducing the detail of questions for the sake of interpretability (i.e. forecasting only binary trends rather than precise numbers).
- Data sources: At once a strong suit and potential weakness of forecasting for policy is the detailed data which should be available to both base predictions on and resolve questions with. In our experience, it is unrealistic to expect comprehensive knowledge of the available data from policy officers. Rather,

available data sources should be discussed with them,[21] and then examined separately to identify potential data points which may be used as references in posing forecasting questions.

5) Day to day administration

Of course, once the partnerships are built and questions developed, they must also be published to be answered by forecasters. While this stage of the process will be very much dependent on the design of your project, there are nevertheless some take-aways we can share from FORPOL that would be applicable to most if not all similar projects.

- There will be typos and other small mistakes that can make it into the final text of the questions or, more frequently, their descriptions - for us, even having three-person peer review, we missed some before publishing.
- The first 24 hours after publishing a question are critical. Forecasters may interpret something in the question wrong, or they may have follow-up questions, and it is important to clear these as fast as possible.
    - It is helpful to have a predetermined process to follow on who is responsible for monitoring these, and if any changes are necessary, who must sign off on these before they may go live.
    - Having a fixed release schedule is very practical also for this reason (beyond helping forecasters adjust their time better etc.) We published new questions on Tuesday evenings, while i.e. GJ Open concentrates question releases on Fridays.
- As a general rule, when running a public tournament it is worth keeping in mind that transparency has costs, and budgeting for them. In other words, communicating and explaining decisions, questions, partner selections - these all take time and attention, but are critical for long-term sustainability. When designing the back-end of the tournament, make sure to allow for buffers which can be used to deal with any matters of this sort as soon as they come up.
- At least in our experience, forecasters tended to be very sensitive to any perceived “restrictions” on their forecasts - use the first few numerical questions to calibrate your own feel of the appropriate bounds/bin sizes[22] compared to the average of your forecasters, and adjust accordingly.

6) Reporting from the tournament

One of our findings which we cannot overstate is the positive feedback to including the rationales of forecasters and weaving a coherent narrative across these by our policy partners. Therefore, we have to reiterate the need to structure the outputs of your

[21] Under the assumption that the tournament is open to the public, this will in most cases be publicly available ones, though the precise availability may vary by country and topic.
[22] As while these would, of course, ideally mostly derive from the policy partners, in our experience this stage tends to become too technical for their involvement and they hand over control to the organisers.

project, no matter its design (one-offs, early warning systems, thematic reports, etc.) to include sufficient space for these rationales.

While we never had to spend significant time convincing our policy partners of the fact that expressing beliefs about future events in terms of probabilities and numerical values is valuable, it is the case that it is not a common type of information to receive normally. Therefore, it can be difficult at times to identify the key messages which i.e. any given probability distribution is communicating beyond the mean and such. Research into the best forms of communication of this information was beyond our scope,[23] but we found some success with prompting closer scrutiny and interest in the data in all its richness by showing alternative readings - i.e. picking a threshold representing an “extreme scenario” and interpreting what its implied probability is.

[23] Though readers may find the discussion in Savelli and Joslyn (2013) interesting, and use it and the references therein as a starting point for more information.