> FHI TECHNICAL REPORT <

Improving Judgments of Existential Risk: Better Forecasts, Questions, Explanations, Policies

Ezra Karger, Pavel Atanasov, & Philip E. Tetlock

Technical Report #2022-1

Improving Judgments of Existential Risk: Better Forecasts, Questions, Explanations, Policies*

Ezra Karger[a] Pavel Atanasov[b] Philip E. Tetlock[c]

[a]*Federal Reserve Bank of Chicago*
[b]*Pytho, LLC*
[c]*University of Pennsylvania*

March 18, 2022

ABSTRACT

Forecasting tournaments are misaligned with the goal of producing actionable forecasts of existential risk, an extreme-stakes domain with slow accuracy feedback and elusive proxies for long-run outcomes. We show how to improve alignment by measuring facets of human judgment that play central roles in policy debates but have long been dismissed as unmeasurable. The key is supplementing traditional objective accuracy metrics with intersubjective metrics that test forecasters' skill at predicting other forecasters' judgments on topics that resist objective scoring, such as long-range scenarios, probativeness of questions, insightfulness of explanations, and impactfulness of risk-mitigation options. We focus on the value of Reciprocal Scoring, an intersubjective method grounded in micro-economic research that challenges top forecasters to predict each other's judgments. Even if cumulative information gains prove modest and are confined to a 1-to-5 year planning horizon, the expected value of lives saved would be massive.

*Our work on this document was funded by members of Founders Pledge. The authors thank Yunzi (Louise) Lu and Josh Rosenberg for extensive support with this research. We also thank seminar participants from the Future of Humanity Institute and the Global Priorities Institute for providing us with feedback on the ideas in this paper. Any views expressed in this paper do not necessarily reflect those of the Federal Reserve Bank of Chicago or the Federal Reserve System.

Table of Contents

INTRODUCTION

This proposal lays out a multi-method plan for generating accurate answers to probative questions bearing on existential risks to our species ("X-risks"). Implementing this plan will be a big job. It will require a new generation of forecasting tournaments that preserve rigorous accountability for short-term accuracy while experimenting with innovations for advancing long-term debates prone to stall out because we lack reliable ways to weigh clashing claims.

Consider one of the gloomiest scenarios that futurists have conjured for the 21st century, one blending several suspected drivers of X-risk (Ord, 2020). Humanity doesn't realize it yet, but we are stumbling along a path toward a Great Power war in the 2030s, a biological-and-nuclear conflagration that, amplified by automated weapons systems, will kill at least 600 million, roughly 10x the death toll of World War II.

First-generation forecasting tournaments could help by doing what they do well: chart the ebb and flow of probability judgments of short-run, objectively verifiable events that are potential precursors of long-run X-risks. But that will only be helpful if the events on our radar screen are actual precursors and imply actionable risk-mitigation strategies (Tetlock, 2017). To reduce reliance on happenstance—and move from prediction to preemption—the next generation of tournaments must explicitly incentivize people to do three things that first-generation tournaments are ill-equipped to do: (a) identify early-warning indicators in a noisy, distraction-laden world; (b) craft insightful explanations that assist policymakers in spotting lead indicators and discounting distractions; (c) give louder voices in policy debates to high-value contributors at each phase of the knowledge-production cycle.

Of course, we are not implying that prediction is easy. Anticipating World War III a decade out would be impressive. But it would be poor form to celebrate anticipating an avertable train-wreck. And moving from prediction to preemption adds formidable challenges. Designers of second-generation tournaments will need to supplement traditional objective accuracy metrics with new sets of intersubjective metrics aimed at fuzzier targets: skill at crafting creative forecasting questions, explanations, and policy solutions.

We address these challenges as well as objections to the entire enterprise. One supposedly killer objection is that we live in a Gray-to-Black-Swan world in which long-range accuracy falls fast to chance—so why invest in foresight? Far from lethal, though, this objection is constructive. It prods us to update expectations about how far it is feasible to see into the future—and explore how timely and accurate forecasts need to be for informing risk-mitigation decisions before "it's too late." Previous forecasting tournaments suggest the heroes of X-risk exercises are less likely to be long-view visionaries than they are short-range belief updaters whose judgments guide timely smaller-scale policy adjustments. Across a surprising range of environments, X-risk tournaments are good bets to deliver value even with moderately myopic forecasters.

TEN CHALLENGES

Challenge 1: Managing Rigor-Relevance Trade-offs. Rigor is not enough; we need relevance, which is elusive because rigorous resolution of the most policy-relevant questions may take decades. By then we may have dallied too long. To inform policy priorities, we need a portfolio of questions with offsetting strengths and weaknesses: some, rigorously resolvable albeit of shaky relevance; others, indisputably relevant albeit of shaky resolvability.

Challenge 2: Seeking Questions That Cleave Nature at its Joints. Recognizing the need for a portfolio and satisfying that need are different things. How can we ensure a flow of candidate questions from which we can pluck the highest incremental-information-value contributors?

Challenge 3: **Understanding Forecasts**. Accurate answers to probative questions count as progress. But stand-alone, numerical probabilities win few hearts and minds; people want explanations that tell them what the numbers mean.

Challenge 4: **From Understanding to Action**. Even numbers accompanied by explanations will not be enough. Policy-makers insist on actionable guidance: What should they do right now? And they aren't eager to sift through reams of metrics to find answers.

Challenge 5: **Recruiting Right Mixes of Talent**. No one knows who will prove most adept at clearing the first four challenges. So, whom should we recruit? How should we blend skills of

generalists (superforecasters) and specialists? What are the right doses of viewpoint diversity?

Challenge 6: Motivating the Talent. X-risk tournaments ask a lot from participants. What is the right mix of intrinsic and extrinsic motivation to sustain individuals and teams for the long haul?

Challenge 7: **Picking Probability Elicitation Tools and Scoring Rules**. Even if we do a good job of picking and motivating people, we won't get the data we need if our scoring rules don't incentivize people to make nuanced probabilistic distinctions (often deep in tail-risk territory) or if our questions don't give them response options for reliably expressing those distinctions.

Challenge 8: Giving Participants Essential Analytic Support. Asking anyone, even seasoned professionals, to leap into X-risk assessments without preparation is itself risky. We should offer training exercises in simulated worlds that give people practice at making micro-probability judgments (below 1%) and at distinguishing stable systems with negative feedback loops from unstable ones with positive loops that could slip out of control, fast.

Challenge 9: Benchmarking. Participants should not become insular. They should compare their performance against a range of external benchmarks, from prediction markets to opinion gurus. How can we help them balance epistemic deference and skepticism?

Challenge 10: Coping with Information Hazards. We need criteria for screening out candidate questions that will help good actors, but not bad ones.

METHODOLOGICAL PREREQUISITES

X-risk challenges raise messy qualitative issues that first-generation tournaments—fixated on objective metrics—push into the background. But X-risk researchers don't have the academic luxury of bypassing awkward questions like "What makes a good forecasting question?" or "What are the best explanations for forecasting failures?" Fortunately, we now know how to render these previously intractable problems at least semi-tractable. The key is developing reliable methods of judging the judgment calls that flow, with minimal-quality control, into policy debates and that first-generation traditionalists tacitly treat as beyond quantification.

Traditionalists err when they define quantification narrowly and restrict tournaments to objective gold-standard accuracy metrics. That restriction rules out a vast set of informative silver-standard intersubjective metrics, that gauge forecasters' skill at predicting other forecasters' judgments on hard-to-score topics: their judgments of the relative plausibility of long-run futures, of the soundness of alternative explanations, and of the forecasting questions most worth posing.[1] To be sure, we understand traditionalists' reservations about soft metrics. Predicting opinions about outcomes does not feel as solid as predicting actual outcomes. So we propose an epistemic compromise. Treat intersubjective metrics of accuracy warily, as imperfect proxies for objective metrics, and confer credibility only when they show statistical value at: (a) reducing noise by tamping down excessive volatility in X-risk estimates over time or random variation at a given time; (b) reducing bias by tamping down systematic variation in estimates; (c) ramping up signal detection by predicting objective indicators of the ground truth.[2]

Our central thesis is that well-constructed intersubjective metrics of X-risk judgment can complement objective metrics and improve high-stakes policy debates faster than would otherwise be possible. We see particular promise in an intersubjective method we call Reciprocal Scoring (Karger et al. 2021), grounded in the microeconomic literature on peer prediction that dates back to Keynes' beauty-contest thought experiment and extends to recent work on Bayesian truth serum.[3] Reciprocal Scoring is not however about predicting the views of ad hoc collections of "peers." It is choosy. It asks carefully culled groups of highly accurate forecasters to predict the views of other

[1] To appreciate the value of intersubjective metrics, it helps to view quantifiability as a continuum, not a dichotomy. At one end are crisp ratio-scale variables, like time. At the other end are fuzzy-set, qualitative categories that defy even rank-ordering, like classifications of literary genres. Between these endpoints is room for creativity, a zone where we can develop intersubjective metrics that let us measure how close and far apart schools of thought are along opinion dimensions: ordinal scales (do they agree x is greater than y and y greater than z?), interval scales (do they see the gap between x and y as roughly equal to that between y and z?), and ratio scales (can we perform multiplicative operations on the perceived values of x, y, and z?).

[2] Our preferred method of separating the drivers of good judgment is Satopaa et al.'s (2021) Bayesian estimation framework: the Bias-Information-Noise (BIN) model. This tripartite division has a long history in statistics and in psychometrics in particular (Kahneman et al. 2021).

[3] We readily acknowledge we are not the first to explore intersubjective metrics. In addition to the Keynesian thought experiment, contemporary research includes work on eliciting beliefs from players in a game where the true state of the world is unknown (Myatt & Wallace, 2012), work on proxy scoring rules (Liu et al., 2020; Witkowski et al., 2017), work on peer prediction (Waggoner and Chen, 2014; Court et al., 2018), and work on Bayesian truth serum (Cvitanic et al., 2019; Frank et al., 2017; McCoy & Prelec, 2017; Prelec, 2004; Prelec et al., 2017).

highly accurate forecasters.

Operationally, Reciprocal Scoring is simple: randomly assign elite forecasters to separate teams tasked with predicting the median of the other team's judgments. Reciprocal Scoring is also flexible. There is no limit on the variety of judging-judgment tasks that teams can do—and then receive rapid accuracy feedback (a prerequisite for learning). Teams can discover in days, not decades, how well they forecast other teams' 25-year forecasts of engineered pandemics—or forecast other teams' judgments of the insightfulness of questions or explanations—judgment calls traditionally written off as hopelessly subjective.

In theory, Reciprocal Scoring should accelerate knowledge production whenever there is latent crowd wisdom: from "what questions should we be asking?" to "what are the most promising policy options?" But the practical value of Reciprocal Scoring hinges on the answer to a game-theoretic question: what is the optimal strategy for forecasters playing under Reciprocal Scoring ground rules?

Our aim is to ensure the optimal strategy in Reciprocal Scoring is no different from the optimal strategy in objective-indicator, first-generation forecasting tournaments: report your best guesses about the true answer. That is the surest path to minimizing proper-scoring-rule functions, like Brier, that gauge gaps between probability judgments and objective reality.[4]

There is though a big obstacle. The truth in first-generation tournaments is objectively resolvable whereas there is no objective truth in Reciprocal Scoring, only intersubjective truth. Overcoming this obstacle is crucial for the scientific credibility of Reciprocal Scoring. Overcoming it will also tax readers' patience because it requires an extended detour linking formal game theory to the nitty-gritty ground rules of Reciprocal Scoring—which we offer in the Technical Appendix.

For readers who are not game theorists, we offer here a short-cut tour that captures the spirit of the formal proof. Game theory tells us that Reciprocal-Scoring forecasts are in equilibrium if no one can earn a higher payoff by unilaterally changing their forecasts. We therefore need to operationalize Reciprocal Scoring so there is only one equilibrium, the desired one in which

[4]We discuss criteria for picking proper scoring rules in X-risk tournaments under Challenge # 7. For deeper exploration of these issues, see Murphy & Winkler (1984) and Clemen & Winkler (1999).

everyone understands the game, carefully studies the questions, and submits their best guesses of what thoughtful forecasters would do. Getting to that point involves setting up ground rules for Reciprocal-Scoring forecasters that eliminate alternative equilibria—such as collusion and low-effort guessing strategies—by rendering them impractical or unattractive.

Once we reach that point, we can put Reciprocal Scoring to work on X-risk challenges: predicting predictions of hard-to-quantify, remote-in-time, objective variables as well as gauges of impossible-to-quantify, intersubjective variables: what counts as a question worth asking, or an explanation worth considering, or a policy worth trying? We can give elite teams of generalists and specialists the task of judging each others' judgments on these questions. When researchers have properly operationalized Reciprocal Scoring, the best move for forecasters will be to invest cognitive effort in proportion to the utility they derive from discovering accurate answers, where that utility will be a function of their intrinsic motivation to reduce X-risks as an end in itself and of their extrinsic motivation to obtain the reputational and monetary rewards that tournaments offer for performance.[5]

Of course, the devil lurks in messy methodological details that lie outside the province of mathematical game theory. The success of X-risk tournaments will ultimately be a function of the cognitive and social support we can give flesh-and-blood human beings struggling to think creatively and rigorously about X-risks—support that needs to take different forms for each of the ten challenges to come.

TACKLING THE CHALLENGES

Challenge 1: Managing Rigor-Relevance Trade-offs. We often have to wait decades for rigorous resolution of long-run questions relevant to X-risk whereas short-run questions about in-

[5]Common sense suggests that intrinsic and extrinsic rewards have additive effects. Revisionists—inspired by Lepper et al. (1973)—maintain that extrinsic rewards undermine intrinsic interest by causing people to over-attribute work effort to material rewards. Counter-revisionists—inspired by Deci and colleagues—respond that extrinsic rewards only undermine intrinsic motivation under a narrow band of conditions: when people see the rewards as signaling the task is aversive or immoral. Our views are closer to Deci and Ryan (2001) who also argue that rewards actually enhance intrinsic motivation when people see them as honoring personal achievement—a position in the spirit of a research program that recognizes "superforecasters."

dicators of X-risks resolve fast but are of debatable relevance.[6] Our two-track strategy for bridging rigor-relevance gaps is to combine objective and intersubjective metrics:

1. Developing clusters of early warning indicators that score high on rigor because each item can be resolved by objective, gold-standard metrics but also score high on relevance because each item has its own distinctive links to long-run risks. The simplest clusters are based on informed observers' hunches about early correlates of a long-run risk (Tetlock, 2017). Consider the "4th Industrial Revolution" scenario which posits rapid advances in AI causing major dislocations in white-collar labor markets by, say, 2040. Figure 1 lays out three early warning signs that experts nominated in 2015 when asked to balance two desirable properties: high correlations with the long-run outcome of interest (maximal relevance) but low correlations with each other (minimal redundancy).

2. Developing silver-standard intersubjective metrics that score high on relevance because they gauge skill at predicting others' views on long-run outcomes we most care about but that still score above zero on rigor because we can validate these metrics against logical coherence and empirical accuracy benchmarks. Intersubjective metrics have two other big advantages: (a) flexibility—we can use them to answer questions that are impractical or impossible to objectively resolve; (b) immediacy—forecasters get rapid accuracy feedback. We don't have to wait until 2030 or 2040 to determine how accurately one group forecast another's predictions on the 4th Industrial Revolution. We find out tomorrow, right after each group has submitted its best guesses about the other group's judgments.

[6]The difficulty of reconciling rigor and relevance drives objections that first-generation tournaments were merely games of trivial pursuits (for a partial rebuttal, Tetlock, 2017).

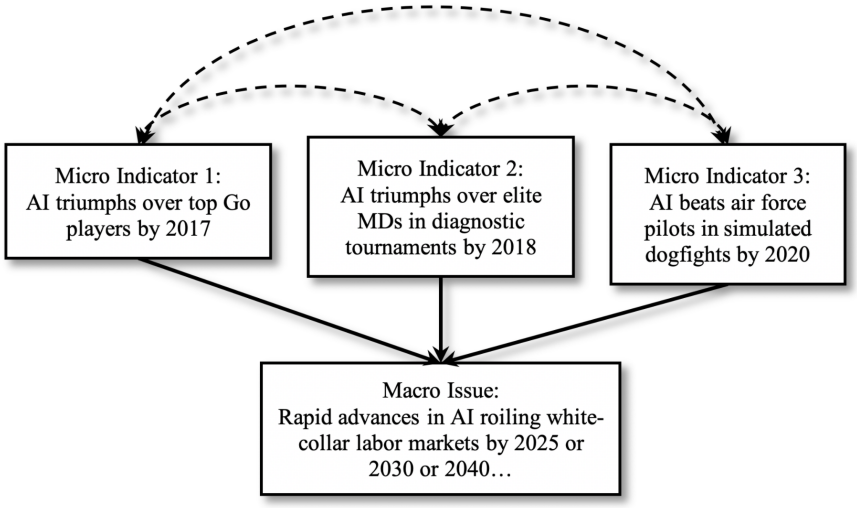

Figure 1: Using objective, short-run micro-indicators to shed light on a long-run macro outcome. Question clusters are correlational and do not require the assumptions about temporal sequencing that more advanced tools require. (See later discussion of conditional trees.)

Gold and silver-standard methods of assessing judgment have complementary strengths and weaknesses. So, invoking the principle of multi-method triangulation, let's explore how each class can compensate for the other's shortcomings, as the need arises.[7]

It arises fast. Application #1 of Reciprocal Scoring is shadowing. Every time we ask forecasters a category-(a) objective question that we will not be able to resolve for years, we also ask them a silver-standard Reciprocal Scoring question that we can resolve almost immediately (by checking the accuracy of predictions of the other team's median forecast). This type of shadowing is always possible because it simply involves pairing objective questions about ground truth X with intersubjective questions about others' predictions of X. Of course, teams will wonder why we are asking what they may see as two variants of the same question: make your best forecasts plus your best forecasts of what the best forecasters will do. Moreover, teams who have read Karger et al. (2021) will know these gold and silver metrics are highly correlated, at least in the short run. Our

[7]Cross-method validation has been a staple of textbooks for decades (Campbell & Fiske, 1959) but underappreciated in work on forecasting tournaments that have been designed by statisticians and micro-economists with a disciplinary affinity for clean metrics (proper scoring rules) and elegant mechanisms (market equilibria). The result has been progress in boosting accuracy but a lack of interest in hard-to-quantify aspects of good judgment, like quality of questions and insightfulness of explanations. As a result, we know less than we should about when objective and intersubjective metrics converge or diverge.

answer to this redundancy objection will be: *"Shadowing serves a vital linkage function in the chain of inference in X-risk tournaments. The core advantage of Reciprocal Scoring is extensionality: its stretchability into domains where objective metrics are elusive. Teams can expect that even when predicting other teams' predictions of events that won't be objectively resolvable for decades, they will still get accuracy feedback within days, even hours. Teams can also expect to be asked to predict other teams' answers to questions that can never be objectively resolved, such as judgments of which forecasting questions will prove most informative or which explanations for forecasts, most compelling, or which policy options, most cost-effective. In all these cases, teams will get rapid accuracy feedback, a prerequisite for learning."*

Our plan is to continually update objective early indicators, like those in Figure 1, so we can gauge how their likelihoods wax or wane over time. Clusters of carefully culled indicators will be more predictable than solo indicators, just by virtue of being less noisy. We also plan to shadow each objective indicator with an intersubjective forecast, creating pairs of variables that Karger et al (2021) have shown correlate highly when they have the same date stamp. The empirical question becomes: how fast will correlations fade when objective and intersubjective indicators have different date stamps—and we test the power of intersubjective forecasters in 2022 to anticipate objective forecasts in 2023 or 2024? Figure 2 projects how we expect the accuracy of both sets of indicators to fade for longer horizons but intersubjective indicators to fade slower than objective ones.[8] The logic underlying this interaction prediction is: (a) long-run objective-indicator forecasters feel less accountable for accuracy than long-run intersubjective-indicator forecasters because the former have to wait 10 to 20 years to find out how well they did and the latter get almost imme-

[8]We view forecasting teams as de facto collaborators to whom we owe full explanations of procedures that might otherwise seem arbitrary. We want them to confront the same epistemic predicament that we, the researchers, confront: the huge gap between what we know and what we need to know about X-risks, a gap that Reciprocal Scoring can partly fill. We will explain the working hypothesis underlying all applications of Reciprocal Scoring: even when objective accuracy metrics are beyond reach, as is often true in X-risk debates, we can create intersubjective metrics that will tamp down noise and bias in their judgment by focusing their analytic efforts on well-defined, proxy-for-truth targets. We will also explain the empirical and game-theoretic rationales for Reciprocal Scoring. On the empirical side, Karger et al. (2021) have shown that asking forecasters to predict each other's predictions often yields predictions as accurate as those we get by asking them to predict objective reality. On the game-theory side, they should understand that Reciprocal Scoring only works if each team assumes the other is playing a purely epistemic game. The unknowability of priors is critical. The logic underlying Reciprocal Scoring breaks down if we don't block lazy shortcuts for coping with accountability.

diate forecasts (a prospect likely to make them more circumspect—Lerner & Tetlock, 1999); (b) past work suggests that when accuracy feedback is delayed, forecasters (especially "hedgehogs") make stronger, theory-driven forecasts (Tetlock, 2005).

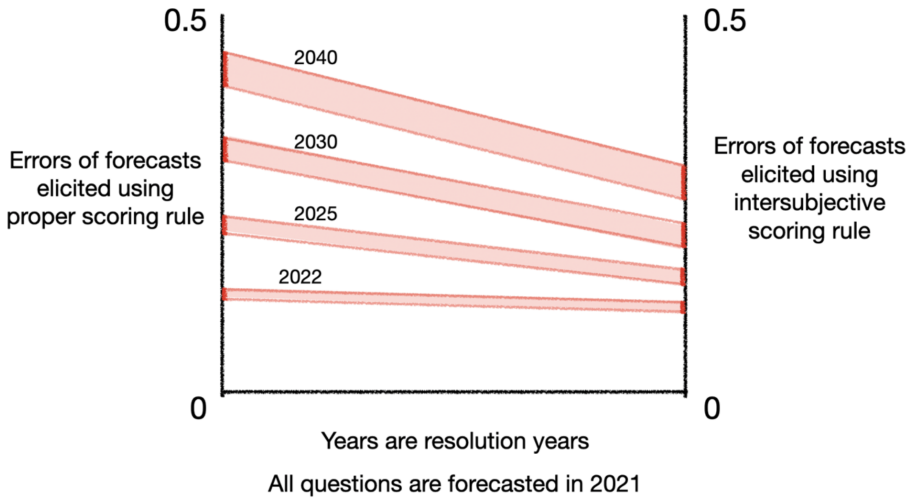

Figure 2: Projected rates of accuracy decay of objective and intersubjective forecasts as a function of time. Projections assume accuracy will not fall all the way to chance for forecasts as far out as 2040 (Brier baseline of 0.5) but will never reach perfect accuracy (zero error), even for short-term forecasts

Let's combine arguments. Figure 1 implies that if we do well at picking objective early indicators one year out—say, 2023—we will also do well at predicting long-run X-risks—say, 2030 or 2040—especially once we know which forecasters got the early questions right and how they did it. Figure 2 implies that if intersubjective Reciprocal Scoring forecasters do well in Application # 1, Shadowing, we will be better positioned to predict, almost right away, how objective early indicators in 2023, 2024,... will resolve. For now, let's arbitrarily stipulate that if these two approaches can jointly boost accuracy by, say, 10%, that counts as our first success in sharpening X-risk foresight beyond existing public forecasting sites.

Of course, all this is hypothetical. No one yet knows how fast predictive links will fade. Moreover, we have no objective method of ensuring the informed observers on whom we rely to pick early warning indicators from the vast universe of possibilities are doing a good job. That puts us back in the same boat as organizers of first-generation tournaments: reliant on subjective judgment

calls about the most informative objective precursors of long-run risks. The solution to Challenge 2 extricates us from that boat.

Challenge 2: Crafting Incisive Forecasting Questions. The question cluster in Figure 1 looks insightful: each early warning indicator has its own distinctive plausible connection to a long-run risk. But face validity is not enough. Sustainable X-risk tournaments must have a rigorous process for generating and selecting candidate questions, which brings us to Application #2 of Reciprocal Scoring.

Question generation is the most cognitively demanding task we will give Reciprocal Scoring teams—so it is prudent to give them practice exercises before generating candidate questions for actual X-risk problems. The exercises will use finite-space games, like Battleship, Goofspiel, and Civ5 where we can measure the efficiency of teams' questioning strategies against algorithmically ideal Bayesian observers. Although teams lose this normative benchmark when they transition into real-world forecasting, the goal of training is to inculcate high-value questioning strategies that travel well across superficially different domains.

Whether in practice or real-world exercises, we will ask teams to organize questioning strategies into conditional trees, which is much harder than assembling question clusters of the sort in Figure 1. Tree construction requires teams to be explicit about their working models of predictive linkages among short-run, micro-indicators of long-run outcomes. A conditional tree takes a final outcome—say, nuclear war—and divides that probability into a sequence of conditioning nodes that posit short-term precursors that have maximal projected impact on the final outcome: trigger events that could escalate an isolated accident into Armageddon. These proximal events should be resolvable within a few years even though the distal event is decades off.

Given that we want to fill forecasting tournaments with high-value conditional-tree questions, we need a means of evaluating trees before questions resolve. To this end, we ask a panel of "superforecasters" to report their best guesses for each question on each branch.[9] The litmus test

[9]We need to formalize the merger process for conditional trees. If two teams' trees are similar, a cross-team dialogue should suffice to iron out differences. If two trees are very different, that is a sign the teams were working with

is the Evidence Ratio. The first branching event, A or not-A (denoted A'), is useful to the extent: (i) it widens variation in probability judgments assigned to subsequent branch forecasts so that the probability of event X conditional on A, p(X|A), departs from p(X|A'), and so on; (ii) it shrinks within-branch variation so tighter consensus emerges on judgments about X, conditional on observing A or not-A.[10] Like the *t*-ratio, the Evidence Ratio is a function of mean-forecast differences between conditional branches (A and A') and pooled cross-forecaster standard deviation (SD) within each branch:

$$Evidence\ Ratio_A = \frac{|Mean(\mathrm{p(X|A)}) - Mean(\mathrm{p(X|A')})|}{\varepsilon + (SD(\mathrm{p(X|A)}) + SD(\mathrm{p(X|A')}))/2} \tag{1}$$

Figure 3 tells a tale of two trees. The superior tree (top panel) cleaves nature at its joints with its first question: how likely are AI systems to pass specific demanding scientific tests of fluid intelligence by date X? "Yes-ish" vs. "no-ish" answers imply different answers to subsequent questions—and the range of views on these questions, conditionalized on answers to the first, shrinks. A yes-ish answer to the first question boosts economic growth estimates on the second question—and this duo of answers elevates estimates of public malaise about AI on the third question, still further out. A pattern emerges: we are heading into a mid-21st-century world where AI is pervasive and threatening. And the chosen nodes separate the high- and low-threat worlds into sub-branches, yielding a high evidence ratio. By contrast, the inferior tree (bottom panel) leads nowhere. It starts with an irrelevancy—will the Red Sox win the 2022 World Series?—a question with zero power to constrain responses to follow-ups or to generate a narrative more coherent than "stuff happens." The evidence ratio for this tree will be close to zero because conditional on the Red Sox's performance, the two sub-branches are as identical as noise permits.

Each Reciprocal Scoring Question-Generation team wants to maximize the Evidence Ratio for

very different theories. At present, aggregating the trees requires some judgment from the researchers; an alternative is to elicit forecasts on both trees. Researchers can later adopt the tree with the better Evidence Ratio. Either way, the stage is set for eliciting objectively resolvable category-(a) forecasts on carefully culled precursor events.

[10]The Evidence Ratio should be used mainly to compare branching events for the same question. The division by the pooled standard deviation of estimates across forecasters functions as a normalization, making Evidence Ratios more comparable across questions.

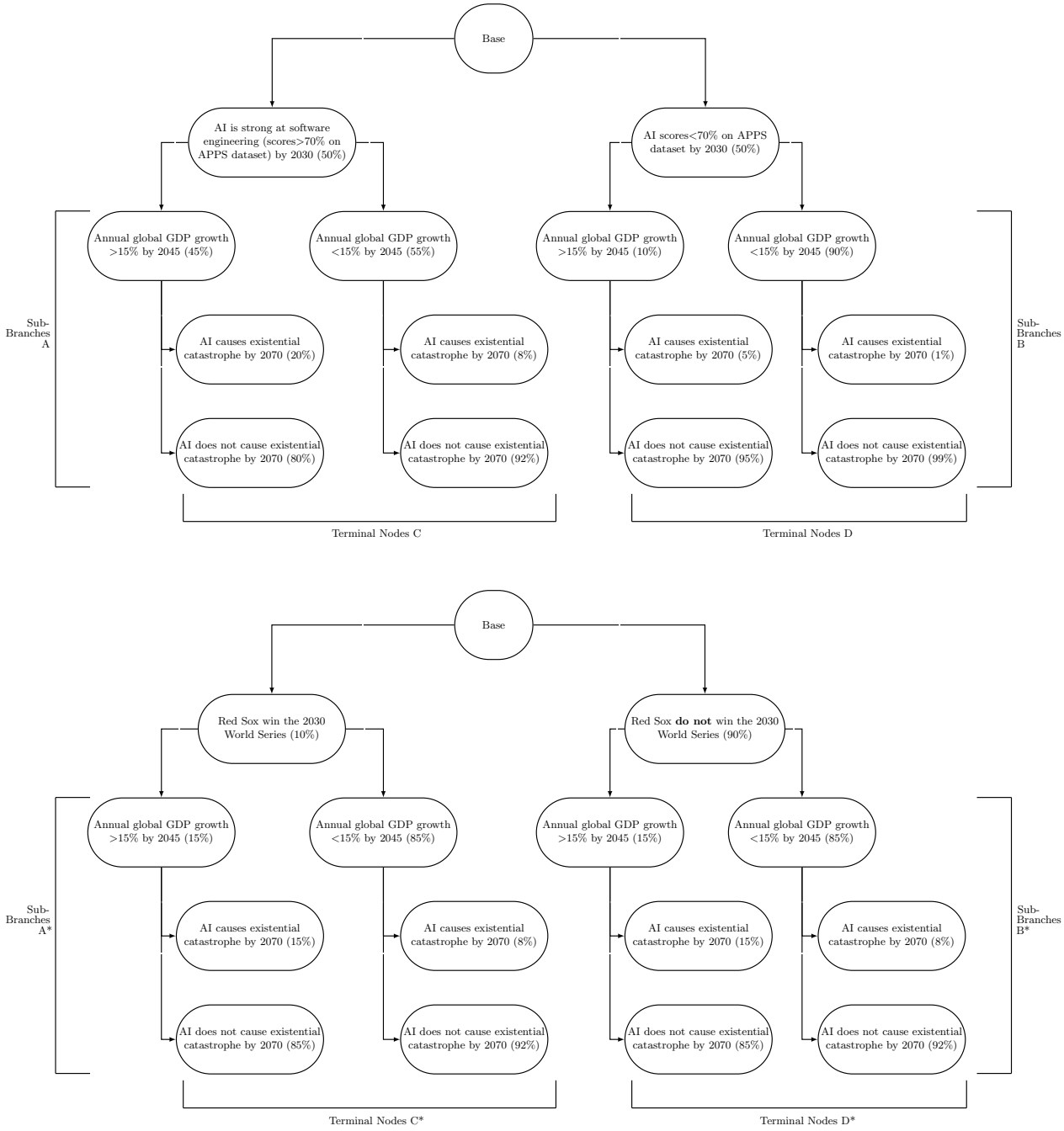

Figure 3: High- and Low-Evidence-Ratio trees appear in the top and bottom panels respectively. Expect the mean gap in probabilities between A and B—and C and D (top tree)—to be (much) greater than the gap between A* and B*–and C* and D* (bottom tree). Expect the variances within A and B—and within C and D—to be (much) smaller than those within A* and B*–and C* and D*

its own ideal conditional tree. Orientation training for these teams will stress that high-Evidence-Ratio trees possess three tricky-to-blend properties: (i) comprehensiveness of coverage so we don't miss key pathways to a given X-risk; (ii) parsimoniousness of coverage so we don't overwhelm forecasters with pathways and trigger events; (iii) incisiveness of questions so the focus is on short-run outcomes with the greatest potential to change minds on long-run risks.[11]

Once a team settles on its tree, it enters the perspective-taking phase: an independent team of highly-accurate forecasters tries to predict each tree, with forecasts on each node elicited using Reciprocal Scoring. Reciprocal Scoring should push each question-generation team toward an equilibrium solution where there is no unilateral incentive to defect to lazy strategies and each team sees its best move as investing cognitive effort in second-guessing themselves: Did we include superfluous branches—or miss significant branches? How should we be making such determinations?

Our hypothesis is that policy-makers will update their views faster in response to X-risk forecasts on high Evidence-Ratio trees, a proposition we can test by running randomized control trials in which policy-makers see data from conditional trees with varying Ratios as well as from ad hoc collections of questions culled from public sites.

Challenge 3: Incentivizing Persuasive, Predictively Powerful Explanations. Rising to the first two challenges advances prediction more than preemption. We can test hypotheses about which schools of thought generate better conditional trees and forecasts. However, another problem looms: probability judgments by themselves, even when nested in easy-to-visualize conditional trees, won't satisfy thoughtful skeptics, even those who believe the answer emerged from sound science. Forecasts without explanations—42%—leave human beings frustrated.

Addressing this challenge requires satiating demand for good explanations. But gold-standard metrics of explanatory quality are elusive. Forecasting accuracy rarely explains itself. One can

[11] In a Bayesian framework, good questions should identify outcomes with likelihood ratios that depart from 1.0 and prior probabilities that rise above zero. So, a Chinese nuclear attack on Tokyo dramatically raises the specter of nuclear extinction but has a minuscule prior probability (say, $p < .00001$). By contrast, naval skirmishes in the South China Sea claiming a few lives barely raise long-run X-risk but are much likelier (say, .01-to-.10). Teams should thus aim for the optimally-informative zone.

be empirically accurate but have a bad theory: be right for the wrong reasons. Or one can be empirically wrong but have a good theory: be wrong yet right on fundamental drivers. Again, the limitations of objective metrics require an intersubjective Reciprocal-Scoring solution. Now though, teams judge each other's explanations. The highest compliment one team can pay another is: "your explanation moved us to change forecasts on which our credibility hinges."

Application # 3 of Reciprocal-Scoring is the Hybrid Forecasting-Persuasion Tournament. Each team's goal is now to nominate explanations that induce the other team to change its forecasts in the right direction. Researchers need to align incentives so teams are equally motivated to avoid false-negative errors of rejecting explanations that would have boosted accuracy and false-positive errors of accepting explanations that would have degraded accuracy.[12]

Accountability to objective metrics is the epistemic backstop for this intersubjective exercise. Team A might win on persuasive impact if its explanations move Team B's forecasts more than vice versa. But the victory is hollow if Team B takes an accuracy hit from heeding Team A's explanations or if A misses a chance to be more accurate because it ignored B's explanations. Such results would be frustrating but also illuminating. They would reveal a perverse inverse relationship between the perceived persuasiveness of explanations and their actual helpfulness in forecasting. Wise teams should treasure these teachable moments.

Winning has a clear definition: develop explanations that help fellow X-risk forecasters. This standard checks temptations to play political games and influence later Risk-Mitigation exercises (e.g., "let's push explanations that exaggerate the acceleration of extreme climate events in order to mobilize support for a policy—carbon taxes—that we know everyone in this elite crowd wants"). Teams will be deterred from such games because they don't want news spreading that the reward for taking their top explanations seriously is degradation of accuracy.

Formally, the persuasiveness of rationale r is a function of the differences in Brier or log scores

[12]We recommend running these hybrid tournaments both before events resolve (Ex Ante Hybrids) and afterward (Ex Post Hybrids). But in Ex Post Hybrids, the challenge for teams shifts from explaining forecasts to explaining their own and the other team's forecasting failures and successes on now known outcomes. How much of a credibility boost or hit should major schools of thought take from better or worse accuracy scores on issue x? And, ultimately, how much should nuclear war rise or fall in the priority queue of X-risks in response to a USA-PRC-Russia arms control treaty—or a North Korean nuclear attack on Seoul or…?

that peer forecasters produced before vs. after exposure to the rationale *r,* across all exposures *n*.

$$Persuasiveness_r = \frac{1}{N}\int_{n=1}^{N}\left[(p_n - y_n)^2 - ((p_n + \Delta_n) - y_n)^2\right] \quad (2)$$

Changes in probability forecasts, Δ_n , can be positive or negative, depending on whether readers of explanations increase or decrease their estimates. Higher values on persuasiveness tell us the rationale moved beliefs *in the correct direction*. Rewards go to educators, not demagogues.

Challenge 4: Incentivizing True Reports About X-Risk Mitigation. Overcoming the third challenge brings us closer to the ultimate goal of preemption. Policy-makers will take us more seriously if they see Hybrid Tournaments shifting elite opinion toward better forecasts and explanations. But policy-makers still want to know what to do *right now,* and won't be eager to sift through reams of metrics for answers.

This sets the stage for Application #4 of Reciprocal Scoring: Risk-Mitigation Tournaments that ask teams of truth-seeking forecasters to predict each other's predictions of policy efficacy. The question becomes: "if we implemented Policy A, B, C, ... immediately (all else constant), how would that change the probability distribution of outcomes from the values in the actual world with no policy interventions?"

Why focus on immediate implementation, an obvious impossibility? Targeting immediate implementation cuts through the ambiguities of conventionally paired conditional questions which blur two issues: How likely is society to do X by date Y? And what will follow from doing or not doing X? A steep carbon tax might be efficacious but appear useless because it can only be implemented after Florida is under water—and we are encased in a runaway greenhouse effect. By spotlighting immediate implementation, forecasters no longer need to ponder political feasibility. They can concentrate on estimating the direct effect of each policy on mortality, assuming implementation, all else equal. By eliciting baseline forecasts about outcomes of interest absent interventions, we can treat the difference between the two forecasts as a direct effect size estimates

of the policy, exactly what policy-makers need to know, right now.[13]

Working independently, each team makes its own policy-conditional forecasts on each option, explains those forecasts, and predicts the other team's policy-conditional forecasts. If teams converge on consistent rankings of policy impacts, the tournament has achieved its goal. If teams diverge, researchers might do another iteration in which teams try to predict how the other team reacted to the first-round forecasts and explanations. We see value here in adapting Delphi-style processes for managing internal team debates: eliciting independent judgments and rationales from team members before entering into group discussions (Landeta, 2006). Teams then make the judgment calls on when their deliberations have reached diminishing marginal returns.

Of course, as with all Reciprocal Scoring silver-standard metrics, cross-team convergence is no guarantee of truth. It may reflect shared misconceptions. Unfortunately, this problem is more severe in Risk-Mitigation Tournaments than in Hybrid Tournaments. The latter are backed by accountability to objective metrics: do explanations boost accuracy? But the former rest on shakier foundations: conjectures about how things would have worked out if we had immediately implemented policy X—counterfactual claims that can never be objectively resolved. As soon as Risk-Mitigation teams predict each other's predictions about immediately implementing a policy —an unrealistic premise—their predictions instantly become inherently counterfactual.

"Inherently counterfactual" need not, however, mean inherently inconclusive. Insofar as Risk-Mitigation teams have access to the best forecasts plucked from Phase-1, High-Evidence-Ratio conditional trees—and access to the best explanations for those forecasts from Phase-2, Hybrid Tournaments—they will have an unusually clear vantage point for identifying cost-effective options. This should raise the chances of cross-team convergence on beneficial policy recommendations, which is why we propose a laborious requirement: teams should predict each other's predictions about the impact of each policy proposal on each branch of the highest Evidence-Ratio

[13]This feature of Reciprocal-Scoring Risk-Mitigation avoids the classic endogeneity problem that arises whenever the conditioning event is more or less likely in better or worse states of the world. It also cuts short intractable counterfactual debates about whether a sound policy undeservedly failed because politicians acted too early or late. And it puts in perspective defenses that forecasters often offer for errant forecasts: my prediction of outcome y would have been on target but for the misimplementation of policy x.

tree. This process will make it easier to pinpoint where teams diverge in their factual-probability assessments and concentrate debate on those points.

Teams should focus on absolute rather than relative risk reduction. Halving the probability of a 1 in 1000 threat is more valuable than halving a 1 in 100,000 threat. Thus, we propose the Absolute Risk Mitigation (ARM) measure: the probability of event X if the proposed mitigation policy (M) is not implemented minus the probability of X if M is implemented. This difference-score approach makes ARMs comparable across policies and risk events.

$$ARM_M = p(X \mid M) - p\left(X \,\middle|\, M'\right) \tag{3}$$

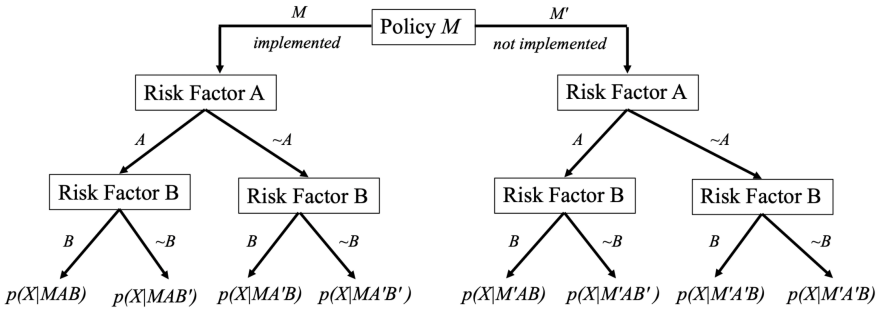

Figure 4: A conditional tree to evaluate effects of policy M on the probability of event X, conditional on risk factors A and B. We can obtain the overall probability of event X conditional on M, *p(X|M)*, by aggregating across branches.

Figure 4 shows how to obtain overall conditional probabilities $p(X|M)$ and $p(X|M')$ from tree decompositions. Calculate $p(X|M)$ and $p(X|M')$ by aggregating estimates from the left and right tree branches, respectively. The aggregation takes into account that each terminal branch probability (e.g., $p(X|MAB)$ is the probability of X given than M is implemented and both risk-factor events A and B have occurred) as well as the probability of reaching each terminal branch.

This process gives harried policy-makers a sounder choice heuristic than business as usual (say, the HIPPO heuristic: follow "Highest Paid Person in the Office"). Mitigation policy M is, by definition, more cost-effective than policy N if Risk-Mitigation forecasting teams conclude that,

per dollar spent, the ARM value for M is higher than that for N.

$$\frac{ARM_M}{Cost_M} > \frac{ARM_N}{Cost_N} \tag{4}$$

Thus far, we have asked teams to focus on estimating policy effect sizes, a variable that policy-makers can then trade-off against competing budgetary priorities. Policy-makers might prefer however to impose a budget cap, B, on risk-mitigation policies, and ask teams to choose the policy i with maximal ARM_i that satisfies the budget constraint. Regardless of this preference, policy-makers can use Reciprocal-Scoring-generated estimates of absolute risk mitigation from each policy to guide their decisions.[14]

Risk-mitigation teams can benefit from grounding their work in all three previous applications of Reciprocal Scoring: in the **best explanations** for the **best forecasts** on the **highest-value** conditional trees. This imposes an empirical discipline that many policy debates lack. Debaters in their haste to win often blur fact and value judgments.[15] Teams should put the burden of proof on members who make strong cost-effectiveness claims but ignore past work products.

These checks and balances may suffice but we recommend an extra safeguard against fact-value conflation. For instance, some risk-mitigation teams might favor the Precautionary Principle and prioritize avoiding errors of commission, which—holding expected value constant—gives an edge to policies already in place. By contrast, other teams might favor classic microeconomic norms of rationality and treat errors of omission and commission symmetrically. Logical-coherence checks can spot when teams diverge because they disagree on facts, values or both.

Imagine two policy options, M and N. In one decision framing, policy M is the status quo.

[14]Imagine three policies– a $1 million policy that reduces risk by 1%, a $100 million policy that reduces risk by 2%, and a $1 billion policy that reduces risk by 5%. Choosing the largest $\frac{ARM_i}{Cost_i}$ leads to adoption of the $1 million policy. But if policy makers can spend up to $100 million on risk mitigation, they should prefer the $100 million policy, even though it is less efficient on a per-dollar basis. The $100 million policy maximizes X-risk reduction given the budget constraint. The $1 billion policy is out of budgetary reach, even though it would reduce X-risk the most in absolute terms. We see value in giving Reciprocal Scoring teams practice in making both pure effect size estimates as well as budget-constrained effect size recommendations. Budget caps are common—and Reciprocal-Scoring teams and policy-makers alike should be sensitized to the choice anomalies that rigid lexicographic decision rules produce (e.g., Tversky's, 1972, elimination-by-aspects model).

[15]Drawing on cognitive dissonance theory, Robert Jervis (1976) called this tendency to conflate fact and value judgments in policy debates "belief system overkill."

Society is already spending $10bn per year on an asteroid monitoring system. In the other framing, spending $10bn per year on the system is just another option on the table. If maximizing lives saved is the goal, we presumably do not want risk-mitigation teams to privilege the status quo. We want them to follow their mandate and redirect all or some of the $10bn to higher-expected-loss threats, say, bioengineered pandemics.[16]

Logical-coherence checks can also spot target non-epistemic distortions traceable to other causes, like variation in moral views on temporal-discount functions (Millner and Heyen, 2021) or on equitable burden-sharing. Of course, we can't promise these cross-checks will yield value-neutral conclusions. There is no view from nowhere (Nagel, 1989). Policy rankings will inevitably be difficult-to-disentangle mixes of fact/probability and value/utility judgments. But the more light we can shed light on otherwise invisible biases, the harder it is for naysayers to dismiss recommendations as politics under a cloak of objectivity.[17]

To recap, clearing the initial four hurdles will give us a transparent process that, relative to business as usual, yields more: (a) accurate forecasts (Application #1 Shadowing); (b) probative questions (Application #2 Conditional Trees); (c) insightful explanations that further boost accuracy and can sway opinion (Application #3 Hybrid Forecast-Persuasion Tournaments); (d) informed estimates of policy impacts (Application #4 Risk-Mitigation Tournaments).

As a rough approximation, Figure 5 estimates the cumulative value-added of all four applications of Reciprocal Scoring by multiplying factors of improvement:

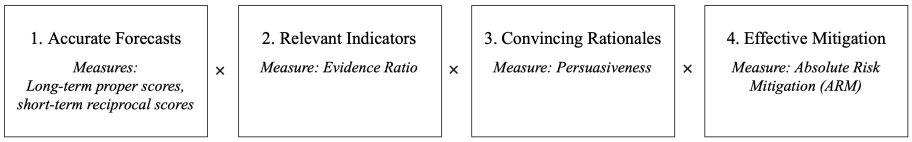

Figure 5: Value-added sources across four applications of Reciprocal Scoring.

[16] One could counter that politics is the art of the possible and coalition-building requires stomaching suboptimal compromises. So we should give Reciprocal Scoring Risk-Mitigation teams an added mandate: adjust their rank-ordering of policy efficacy in response to political feasibility.

[17] Just as we can run both Ex Ante and Ex Post Hybrid Tournaments, we can do the same for Risk-Mitigation. How well do rankings of Risk-Mitigation proposals hold up in hindsight? Ex Post Risk-Mitigation Tournaments are disciplined policy post-mortems that ask each team to offer best guesses of what they and the other team got right or wrong. The goal of these Ex Post exercises is to help Ex Ante teams see themselves from the outside—and reduce judgmental distortions like hindsight bias, ego-defensiveness ...

To define baseline performance, suppose a policy-maker is worried about biosecurity risk, which unbiased forecasters believe has a 10% chance of wiping out humanity by 2100. The policy-maker asks a trusted expert to choose one risk mitigation proposal, and this standard advisory input, with no methodological innovations, reduces risk from 10% to 8%.

Now let's add the innovations. Working backward through Figure 5 and starting with Application #4, we divide top biosecurity experts into two Reciprocal Scoring teams that predict each other's estimates of policy impacts. We hypothesize this exercise will lead to a policy that reduces risk an additional 20%: a 0.02*__1.2__ = 2.4% cumulative reduction of absolute risk.

We also expect that experts in Application # 4 will do a better job if informed by the output of the previous three Applications:

- Application #3 (the most persuasive rationales for forecasts, a reduction of absolute risk by a factor of 1.25: a cumulative absolute risk reduction of 0.02*1.2*1.25 = 3%;
- Application #2 (the highest Evidence-Ratio conditional trees, an additional 10% reduction in absolute risk: a cumulative reduction of 0.02*1.2*1.25*1.1 = 3.3%);
- Application # 1 (the best forecasts of objective and intersubjective indicators, an additional 30% reduction in absolute risk: a cumulative effect of 0.02*1.2*1.25*1.1*1.3 = 4.3%).

Total improvement across phases of knowledge production is: 0.02*1.2*1.25*1.10*1.3 = 4.3%, reducing the original 10% risk of extinction to 5.7% (against the business-as-usual scenario of 8%). We call this combined benefit **information gain**. And if the projected effect sizes look modest, we recommend putting them in an expected value perspective: a 2.3% improvement in our chances of saving a billion lives translates into 23 million expected lives saved.

Of course, these effect size estimates are guesswork now.[18] And even if we skillfully implement each application of Reciprocal Scoring, process innovations alone won't get us over the

[18]One could argue that the each application's value is additive and independent. But the multiplicative approach has two advantages: (a) it captures the compounding returns from investing in methodology; (b) it pushes us to steer clear of value-subtracting applications. If botched construction of conditional trees reduces relevance by 30% (rather than raising it by 20%), we suffer a cumulative information loss even if all other factors behave as noted above.

remaining hurdles. We still need to recruit the right people and give them the right support. The next challenges are about balancing forecasting talent and viewpoint diversity in assembling teams; preparing teams for distinctive features of X-risk problems; identifying external-to-the-tournament benchmarks for assessing how well tournament forecasters are doing; and managing the information hazards that may arise in inquiries into X-risks.

Challenge 5: Recruiting the Right Talent—and Committing to the Mission. No one knows who—which mix of humans, AI and human-AI hybrids—will prove most adept at forecasting looming X-risks. But we can make educated guesses today about how to squeeze the most signal value from human judgment. And we prioritize resolving two debates which we simplify here.

One pits cognitive egalitarians who bet on viewpoint diversity (Page, 2006)[19] against elitists who bet on spotting talent ("superforecasters," Tetlock & Gardner, 2015).[20] An ideal research design would contrast forecasting teams that maximize viewpoint diversity within skill constraints vs. teams that maximize skill within diversity constraints. A second debate pits defenders of traditional prestige hierarchies of expertise against disruptive challengers who embrace generalists with strong predictive records. Here an ideal design would contrast forecasting teams that maximize predictive track records within domain-knowledge constraints vs. teams that maximize domain knowledge within predictive-accuracy constraints.

Based on past work, our best bet is that maximizing forecasting skill within cognitive-diversity constraints will beat maximizing cognitive-diversity within forecasting skill constraints on the dependent variable of forecasting accuracy. But when we switch to dependent variables like incisiveness of conditional trees and persuasiveness of explanations, the advantage may well switch to the strategy of maximizing diversity within skill constraints. We see it as an even-money bet

[19] See Scott Page's (2004, 2006) diversity theorem, which in its simplest form declares diversity trumps ability.

[20] See Philip Tetlock and Dan Gardner (2015) on the methods and superforecasters behind the success of the Good Judgment Project in forecasting tournaments. The diversity-vs.-talent dichotomy is misleading here because talented forecasters tend to internalize diverse perspectives and have animated conversations with themselves (Karvestki et al., 2021; Tetlock, 2005). Talent spotting without ground-truth accuracy is a challenge; Atanasov & Himmelstein (2022) show that intersubjective and behavioral measures also provide strong talent-spotting signals.

whether mixed generalist-specialist teams will be able to achieve cognitive dominance and outperform pure-specialist and pure-generalist teams across the dependent variables of forecasting accuracy, conditional trees and explanations. Finally, although we expect superforecaster generalists to out-predict specialists, we suspect the latter to have an edge in developing conditional trees of warning indicators that empirically dominate (in terms of Evidence Ratios) the proposed conditional trees of superforecaster generalists. Though that advantage will be offset by the tendency of specialists to make more false-positive errors in predicting events consistent with their theoretical commitments and to be slower to notice when their conditional trees begin to deviate from those priors.

Challenge 6: Motivating the talent. Whatever the mix of personnel, we need an organizational culture that elevates the pursuit of truth to the status of *transcendent mission* and aligns norms for teamwork accordingly (Tetlock & Gardner, 2015, on perpetual beta cultures). The ideal personality profile for this mission includes: (a) a burning desire to preserve sentient life in our tiny sector of the galaxy as humanity passes through a precarious technology-transition period; (b) a monastic commitment to the superforecaster agenda of minimizing gaps between their individual judgments and reality—and of separating fact from value judgments.

Of course, few fit this profile. So we should pre-test varying blends of intrinsic and extrinsic motivators for nudging individuals and teams in the desired direction: either via symbolic recognition as "superforecasters," "super-question-generators" or "super-persuaders" or via material rewards for producing high-value conditional trees, forecasts, explanations, or estimates of policy impacts.

We should also ensure the social atmospherics of X-risk tournaments honor those who do things that push the epistemic mission forward, especially at those junctures where the knowledge production process is most vulnerable. We suspect that generating high-value conditional trees of questions will severely tax the patience of Reciprocal Scoring teams—as will the initial solo-work requirement that all forecasters tackle a minimum number of questions on their own (both

probability estimation and explanation of estimates). We have also found that forecasting teams, even superforecasters, struggle to implement best group-deliberation practices for checking excessive conformity in teams (standard checks are devil's advocacy and red teaming) and for resolving deep disagreements (standard check is adversarial collaboration). Jumpstarting self-correcting epistemic communities is a nontrivial task.

Challenge 7: Picking Probability-Elicitation Tools and Scoring Rules. Even if we plug in the right mix of forecasting talent and viewpoint diversity into the right organizational culture and incentive matrix, mistakes will be inevitable. Even well-behaved time-series of data points are occasionally punctuated by abrupt discontinuities of the Gray-to-Black Swan sort: say, an exponential surge in capabilities of fusion reactors or quantum computers.[21] Researchers have long known about the instability of tiny probability estimates (Kahneman & Tversky, 1979), that people are prone to over-weight close-to-zero probabilities that questioning makes salient and underweight, even ignore, them otherwise—one of the cleanest examples of a Heisenberg effect that experimental psychology has to offer.

We propose a three-part solution to these measurement challenges:[22]

(1) Adopting Karger et al.'s (2021a) non-parametric probability elicitation tool, which frees forecasters from researcher-imposed, binned-response options and lets them choose cut-off points along an outcome continuum that correspond to their best guesses of specified likelihoods. For instance, how low would deaths (due to cause X) have to fall to be considered less than 50% or 25% or 1% or 0.1%...likely?;

(2) Asking forecasters to use conditional-tree diagrams that are designed to slow the descent of compound events into micro-probability ranges and that make explicit the hypothesized

[21] In 2013, Tetlock attempted an adversarial collaboration with Taleb in a joint SSRN paper that they never completed but remains relevant. The collaboration collapsed partly because Tetlock did not see "anti-fragility" as a viable alternative to probability estimation. Anti-fragilizing institutions against extreme tail risks is extremely expensive, which requires setting priorities, which in turn requires estimating probabilities. That said, it is easy to imagine probabilistic variants of the Precautionary Principle that teams might invoke in risk-mitigation tournaments.

[22] See Daniel Kahneman & Amos Tversky's groundbreaking work on prospect theory (1979) and its refinement, cumulative prospect theory (1992), especially probability weighting (pi) function descriptions.

causal contingencies underlying probabilities;

(3) Switching to log-based proper scoring rules that are more sensitive than quadratic Brier scoring to variations in accuracy in low-probability ranges and align incentives more sensibly with mistake-avoidance priorities in the X-risk domain. Brier-scored forecasters can still recover from the maximum credibility hit they take from declaring something impossible (zero probability) that then happens or declaring something certain that then does not happen. By contrast, log-scored forecasters never recover. The penalty for wrong-headed absolutist thinking is a negative-infinity credibility score. In the scoring rules below, f_c denotes the estimated probability for the correct outcome in a binary question:

Brier Scoring Rule = $2 \times [1\text{-}f_c]^2$,

$$\text{from } 2 \text{ (worst) to } 0 \text{ (best)} \tag{5}$$

Logarithmic Scoring Rule=$\ln(f_c)$,

$$\text{from } -\infty \text{ (worst) to } 0 \text{ (best)} \tag{6}$$

Note that this three-part solution does not tamper with a methodological principle that forecasting researchers take as axiomatic: tournaments should be governed by proper scoring rules that incentivize forecasters to report their true beliefs—and treat errors of under- and over-estimation symmetrically. It is however reasonable to wonder whether proper scoring rules are appropriate for X-risk tournaments. Some scholars argue "no:" under-estimating X-risks is (much) more serious error than over-estimating them and our scoring rules should reflect that asymmetry. Others counter that embracing improper scoring rules would be profoundly improper. It would undercut our long-run scientific credibility.

This dispute leads to testable predictions. Schools of thought that see X-risk threats as imminent (and concerns about long-run credibility as foolish) will be most supportive of: (a) over-

estimating risks; (b) denying the over-estimates are strategic and insisting they represent true beliefs (deception is justified if the stakes are high enough); (c) replacing value-neutral, proper scoring rules with value-skewed rules that treat false-negative X-risk forecasts as far more serious than false-positive ones.

Our view is that modifying proper scoring rules would confuse (and exasperate) forecasters and the right juncture to introduce asymmetric utility functions is after we submit tournament-generated unbiased forecasts to policy-makers but before policy-makers settle on X-risk pre-emption priorities. If policy-makers will only heed the risk to a billion lives if the probability is above a "noticeability threshold," say 1%, the research community should focus on educating the policy-makers, not on distorting the forecasts, which we see as a dangerously credibility-corrosive game. Proper scoring rules help preserve a methodological firewall between accuracy and advocacy objectives.

Challenge 8: Helping People Prepare for Distinctive Analytic Challenges of X-Risk Assessment. Mastering advanced probability-elicitation tools doesn't come naturally even to sophisticated generalists and specialists. Nor do many of the activities noted under earlier challenges, like constructing high-value conditional trees and making logically consistent forecasts across varying time horizons.

For these reasons, we plan to give participants, at the outset, practice exercises in simulated worlds where they can acclimate. Ideal simulations capture the views of competing schools of thought on key functional features of real-world problems, such as figuring out the interplay of negative or positive feedback loops or of assessing probability distributions of possible worlds activated by policy interventions. For instance, forecasters could learn to master climate models that imply little potential for deviations from long-term trends as well as models that imply exponential change—with parallel exercises in other content domains—e.g., pandemics, nuclear deterrence—to promote the transfer of training.[23]

[23]We are not implying this will be easy. A century of work has shown how hard it is to achieve large replicable transfer-of training effects across even superficially different content domains (Simon & Hayes, 1976; Gick &

Simulations should also give participants practice at doing reflective-equilibrium thought experiments, with an emphasis on avoiding sub-additivity traps (Tversky & Koehler, 1994), in which they wind up putting a lower probability on a set of outcomes than on its exclusive and exhaustive subsets.[24] Logical-coherence checks on forecasts are especially valuable when empirical accuracy indicators are elusive and even careful judges can be erratic. Both of these conditions are satisfied by extreme events with judged probabilities close to zero.

We recognize that Challenge 8 constitutes an ambitious research program in its own right, more ambitious than IARPA's FOCUS program on counterfactual forecasting which cost tens of millions of dollars from 2018 to 2021. The goal is to enhance cognitive ambidexterity: rapid adaptation to worlds that do or don't mesh with one's priors.[25]

Challenge 9: Benchmarking Against External Standards—and Balancing Deference and Skepticism. We tend to fall in love with our inventions. And there is no shortage of inventors touting tools for boosting foresight—and no guarantee the methods developed for X-risk tournaments will out-perform simpler alternatives powered by less expensive talent. That is why we will encourage participants to adopt a stance of intellectual humility—and compare their performance against external benchmarks, including opinions of prominent experts (e.g., Ord, 2020),[26] prediction markets,[27] and relevant statistical models. We will also remind participants of how hard it has repeatedly proven for humans to beat even simple extrapolation algorithms (Dawes, 1979;

Holyoak, 1983; Gentner, 1983; Kotovsky, Hayes, & Simon, 1985; Gentner & Markman, 1997).

[24] An example of temporal-scope insensitivity would be judging (p (war next 3 years $<$ p (war) in each of years 1, 2 and 3) and an example of geographic-scope insensitivity would be judging (p (war anywhere) $<$ p (war in each regional theater). Tversky and Koehler's (1994) support theory predicts such patterns of inconsistency in probability estimation across alternative framings of the same problem.

[25] Cognitive ambidexterity can be defined in many ways beyond the brief definition here (Patil & Tetlock, 2014).

[26] Toby Ord's 2020 book features a comprehensive summary of existential risks for the next century, with probability estimates (e.g., Chapter 6), which we will use as a source of external benchmarks for forecasters.

[27] Prediction polls elicit probability estimates from forecasters working independently or in groups, provide accuracy feedback in the form of proper scores and aggregate estimates statistically. Prediction markets elicit estimates as bets and forecasters (traders) compete for play- or real-money. Prediction markets are an alternative to prediction polls and have advantages, such as continuous intermediate feedback (oscillating prices). We see however three reasons to prefer prediction polls as the major elicitation mechanism for X-Risk tournaments: (a) polls tend to beat markets on longer-duration questions: months or years vs. days or weeks, as shown by Atanasov et al. 2017; (b) polls permit more flexibility in accuracy scoring and feedback provision; (c) polls are better at spotting skilled forecasters.

Kahneman, 2011; Meehl, 1954; Tetlock, 2017).[28]

When big gaps arise between tournament-generated probabilities and those of external sources, we recommend doing what "superforecasters" are supposed to do in their internal team debates: strike sensible compromises between deference (what do they know we don't?) and skepticism (what do we know they don't?)—and reach out to the other side to set up adversarial collaborations in which the parties agree, ex ante, on the evidentiary conditions for changing their minds (Kahneman, 2011). If participants balk at adversarial collaborations, researchers face a judgment call: whether to launch a fifth application of Reciprocal Scoring aimed at improving peripheral-vision, an exercise we call Blindspot Spotting. Each team tries to anticipate the other team's best guesses of where X-risk forecasters are at most risk of being blindsided.

At this juncture, we realize we have already woven second-guessing into every phase of the knowledge-production process so it is reasonable to wonder: Will more second-guessing improve or degrade judgment? Either way, however, it should be feasible to identify underlying mediators, using tools such as Satopaa et al.'s (2021) Bayesian Bias-Information-Noise model. Does Blindspot Spotting sharpen signal detection or degrade it by introducing biases (e.g., pumping up probabilities of scenarios that second-guessers made salient) and noise (e.g., inducing excessive volatility by calling attention to tangential irrelevancies)?

We don't have the answers yet but we do have hypotheses. When stakes and uncertainties are at X-risk magnitude, people will often fall back on well-rehearsed cognitive routines, such as more fox- or hedgehog-like modes of thinking (Berlin, 1950; Tetlock, 2005). The prototypic fox enjoys intense second-guessing whereas the prototypic hedgehog grows weary fast. Such tensions will rise and fall as hedgehogs, besieged by dissonant data, make grudging concessions to complexity and foxes, overwhelmed by diversity, retrench and simplify.

Challenge 10: Managing Information Hazards. We left the toughest obstacle for last for it becomes problematic only if we pass the previous challenges so convincingly that the fruits

[28] See the second edition of Philip Tetlock's (2017) Expert Political Judgment—and the description of expert accuracy vs. algorithms of varying sophistication in Chapter 2, The Ego-deflating Challenge of Radical Skepticism. Also see Daniel Kahneman's (2011) summary of this work in Chapter 22: Expert Intuition: When Can We Trust It?

of our labor—questions, answers, explanations, policy rankings— command wide attention. The question now becomes: Are our findings more helpful to good actors trying to prevent catastrophes or to bad ones trying to trigger them?

Bostrom (2019) makes a powerful case that humanity is passing through an exceptionally vulnerable era in which accelerating technological advances are outpacing our capacity to contain devastating side effects.[29] So, there are strong reasons for caution. It would, for instance, be reckless to sponsor a contest in which experts compete to predict each other's predictions of which pathogens could be most efficiently gene-edited to wreak maximal havoc. Perhaps intelligence agencies should consider such contests under veils of secrecy but, again, only if the benefits of developing countermeasures out-weigh the risks of leakage to bad actors.

What criteria should researchers use to draw boundaries between permissible and impermissible knowledge of X-risks? We see ourselves near the midpoint of an opinion continuum anchored by opposing laissez-faire and dirigiste perspectives.

At one end of the continuum is the laissez-faire camp that chafes at restrictions on freedom of inquiry. It is on a well-defined, neo-positivist, signal detection mission: clarify key unknowns by (a) estimating probability distributions of the effects of false-positive mistakes (allowing studies that help bad actors more than good ones) and of false-negative mistakes (squelching studies that help good actors more than bad); (b) exploring where observers with varying priors and risk preferences set thresholds for deciding how high risk/benefit ratios must be to squelch studies.

At the other end of the continuum, the dirigiste camp embraces expansive versions of the Precautionary Principle and vetoes research whenever a credible epistemic authority can spin a plausible scenario about how bad actors could use empirical findings to advance nefarious ends.

We are wary of both camps. Pursuing the signal-detection program will let risk/benefit ratios rise far above 1.0. Embracing the Precautionary Principle will ensure that numerators in risk/benefit ratios never rise above zero. To invoke a maxim of cybernetic learning theory, it is hard to define stopping rules that prevent both over-shooting and under-shooting—that stop us

[29] Bostrom, N. (2019). The vulnerable world hypothesis. Global Policy, 10, 455-477.

right after we have gotten enough and before we get more than enough.

Our approach to Information Hazard debates parallels our approach to Risk-Mitigation: help debaters disentangle the influence of fact/probability and value/utility judgments on policy recommendations. Imagine two factions, each with the same value function for humanity's future. But Faction #1 leans laissez-faire because it sees only a minuscule probability of advances in destructive capabilities out-pacing advances in defensive measures. Benefits exceed losses by hefty margins in 999 of 1000 reruns of history with unregulated debate—though the losses in the 1 out of 1000 cases are catastrophic. Faction #2 leans dirigiste because it holds the mirror-image view, a 999/1000 chance of offense trumping defense. They see catastrophe as almost a sure thing if the laissez-faire faction prevails. And the laissez-faire camp sees massive opportunity costs as almost a sure thing if the dirigiste faction prevails.

In theory, these two factions should converge in their factual-probabilistic assessments if they participated in an Information Hazard Reciprocal Scoring exercise, which would be a sixth application of the technique. We could ask each camp to predict the other's predictions of the judgments of an elite, truth-seeking panel of scientists (a panel that need not exist but ideally would). We could then document how much the probability of the pandemic-offense-dominance hypothesis would have to rise for Faction 1 to abandon its laissez-faire position (say, from 1/1000 to 10/1000) and how much the probability would have to fall for Faction #2 to abandon the strong Precautionary Principle (say, from 999/1000 to 10/1000). In this stylized example, the factions have the same value/utility function so they converge on the same regulatory-policy stance as soon as they converge on beliefs.

Actual debates will be more complex and it is psychologically implausible that two factions with starkly different views on facts would have identical value/utility functions. But long journeys begin with initial steps. Systematic policy-capturing research in the spirit of Hammond and Adelman (1976) can partition disagreements among schools of thought into disputes over facts and over values. And for disputes grounded in facts, adversarial collaboration in the spirit of Kahneman (2011) has had several successes in inducing feuding camps to make testable predictions that

shed light on the merits of their positions.

Risk regulators face a tricky choice: between covert/impressionistic/subjective vs. overt/systematic /intersubjective methods of assessing the safety of accelerating learning in X-risk domains. We prefer the latter approach but readily concede our lack of substantive expertise. Our preference—an open inquiry into how open we should be—could be too risky in, say, the bioterrorism domain. Our working philosophy is therefore this: Slow down when top scientists flash amber or red lights, as they now apparently are for engineered pandemics. Otherwise, full speed ahead.[30]

CLOSING THOUGHTS: ESTIMATING IMPACT ON X-RISK POLICY DEBATES

We see scientific value in combining objective and intersubjective metrics: a new generation of tournaments with greater policy relevance than first-generation tournaments but at little cost in rigor. Of course, we could be wrong. It is reasonable to ask us to do a pre-mortem: to imagine possible causes of failure and specify how we plan to preempt them.

The obvious objection is that we are too optimistic about raising predictive accuracy above chance or, harder still, extrapolation algorithms. We inhabit a noisy, Gray-to-Black-Swan world in

[30]Bostrom (2019) advances a provocative thought experiment in which, prior to World War II, physicists discover a stunning shortcut to manufacturing fission bombs, one easily within the technical reach of any nation-state. This hypothetical initially looks like a slam-dunk demonstration of the need to exercise self-restraint in exploring ideas with mega-destructive potential. A nuclearized Nazi Germany is a nightmare. But the rules of this thought experiment also imply a nuclearized UK, USA, France, USSR and Poland. Would Hitler have invaded Poland if he knew that the Poles could quickly obliterate his advancing panzer columns, not to mention German cities. A counterfactual controversy looms: Does the Bostrom hypothetical gate us into runs of history in which World War II claims 600 million lives, not just the 60 million we suffered? Or into runs of history in which a nuclear-deterrence balance of terror prevents World War II, and saves millions? No one knows for sure but the diverging forecasts remind us of a debate thirty years ago among political scientists on the robustness of nuclear deterrence (Sagan & Waltz, 1995). Waltz argued that nuclear proliferation actually makes the world safer—a position that Sagan argued rested on unrealistic assumptions about human rationality and the infallibility of nuclear command and control. In our view, the Bostrom hypothetical implies a pattern of nuclear hyper-proliferation to sub-national actors that strains the Waltzian interpretation to the breaking point. But the diverging forecasts do still underscore reasons for caution in arguing for scientific censorship: (a) even clear cases for self-restraint can become murky under close inspection; (b) the murkiness will be a function of the mental models of causality we import to fill in imaginary counterfactual probability distributions of deaths (Tetlock & Belkin, 1996). If we were to role-play counterfactual historians, we would bet on the tail risk death toll of rapid proliferation prior to World War II exceeding the possible benefit of averting World War II. So Bostrom's argument holds. But let's remember the speculative nature of the exercise—and the value of Reciprocal Scoring for reining in excessive speculation in debates of this sort.

which long-range accuracy falls fast. So why invest in foresight? In this view, we should spend the money on "anti-fragilizing" institutions against big shocks that lurk so deep in the tails of probability distributions of possible worlds that it is futile to estimate them.[31]

To explore this objection, imagine collecting objective and intersubjective forecasts every year, starting in 2022, for outcomes up to 2072. By 2025, we are positioned to answer questions that shed some light on the value of longer-run forecasts: (a) How fast does accuracy fade along the 1-to-3 year time-horizon?; (b) Do aggregate time-series of forecasts about 2030 exhibit excess volatility as, say, defined by the Augeblick-Rabin criterion?; (c) Are individual forecasters updating in small or large increments?;[32] (d) Do we observe large cumulative updates between 2022 and 2025 for the 2030 outcomes? Suppose the answers are always yes, so we have grounds to be bearish about the accuracy of 2022 forecasts aimed at 2030—not to mention 2050.

Should we then write off X-risk tournaments? Our answer is no. Here we draw on econometric modelers who treat the connections between forecasting and policy-making as a stochastic-dynamic control problem (Millner & Heyen, 2021). Across a surprisingly broad band of simulated conditions, they find that policy-makers can, at acceptable cost, substitute long-run forecasts with short-run ones as long as they retain a capacity to adjust decisions in response to new data. Athey, Chetty, Imbens, and Kang (2019) put a similar idea into practice when they estimated the long-run effects of childhood interventions on adult outcomes using short-run surrogate indices correlated with long-run outcomes. To condense what would otherwise be a lengthy digression, our emphasis on short-run proxies for long-run events draws on a large body of theoretical and applied work in economics and finance (Harvey, Rattry & van Hemert, 2021).

We hypothesize the policy usefulness of tournaments to be a joint function of the accuracy of the forecasts it generates and the rapidity with which it delivers accurate forecasts.

$$Policy\ Utility = Accuracy \times Rapidity\ of\ Delivery \tag{7}$$

[31] As noted elsewhere, anti-fragilizing is expensive–and requires setting priorities which requires at least tacit probability judgments.The alternative to relying on explicit estimates of uncertainty is relying on implicit estimates.

[32] For a discussion of how to measure excess volatility, see Augeblick & Rabin (2021). For a discussion of the predictive value of patterns of individual updating, see Atanasov et al. (2020).

Usefulness depends on accuracy because accurate forecasts help us spot superior policies (relative to a baseline counterfactual world without such forecasts).[33] We define accuracy as improvement in proper scores above that of ignorance priors and we assume accuracy declines with temporal distance.[34] Usefulness also depends on rapidity of delivery because the sooner we get accurate forecasts to policy-makers, the faster they can adopt sensible risk-reduction policies. We normalize and bound values of accuracy and rapidity of delivery to the [0,1] range. So, a zero-accuracy forecast has zero usefulness even if we guarantee instantaneous delivery—and a perfectly accurate forecast has zero usefulness if it arrives too late.

We hypothesize that extending time-horizons has opposing effects on forecasting accuracy and rapidity of delivery. Figure 6 lays out our best guesses. As we stretch horizons from 2 to 10 to 50 years, expected accuracy of human judgment falls toward chance.[35] Although the research literature is less helpful than ideal, it suggests that accuracy degrades at a non-constant rate (Himmelstein & Budescu, 2021): faster initially, then flattening as resolution dates approach. As a rough approximation, we model this non-linearity as an exponential reduction in accuracy over time (square-root of years to resolution). [36]

We split Figure 6 into left and right panels because we expect faster accuracy decay when forecasters feel less imminently accountable for their judgments. In the earliest geopolitical tournaments, Tetlock (2005) found the fastest accuracy decay among "hedgehogs making long-run (5-year) forecasts inside their domains of expertise"—an effect he attributed to delayed feedback weakening accuracy accountability and thus emboldening forecasters to engage in theory-driven

[33] A comprehensive formulation would include question relevance and persuasiveness of rationales. We use the simplified formulation that omits these variables to focus on the effects of time-horizon on forecast usefulness.

[34] Tarsney (2019) sees a pattern of deteriorating forecasting accuracy for longer time horizons across fields—and views that pattern as an existential threat to longtermism. Focusing on the farther future may also make it harder to generate probative questions and persuasive rationales. To the extent that relevance and persuasiveness also degrade over time, that further strengthens the case for relying on shorter-run proxy questions.

[35] First-generation tournaments used Brier scores of 0.5 as the point where accuracy reaches zero on a binary-choice forecast. In the X-risk context, however, extreme-event probabilities, the use of logarithmic scoring and more sophisticated benchmarks make defining the zero-bound for accuracy skill scores more challenging.

[36] The precise functional forms and parameters for Expected Long-Term Accuracy and Rapidity of Delivery are: $\textit{Long Term Accuracy} = \textit{Short Term Accuracy} \times d^{\sqrt{\textit{Years to Resolution}}}$ $\textit{Rapidity of Delivery} = \textit{Log}\left(\sqrt{1+\textit{Years to Resolution}}\right) + 0.1$ In the Long-Term Accuracy formulation, d is the exponential decay constant that we hypothesize will be raised (improved) by applying intersubjective scoring measures.

(as opposed to data-driven) thinking about possible futures. If true, accuracy should decline more slowly among Reciprocal-Scoring forecasters making long-run judgments than among objectively scored forecasters. As noted earlier, Reciprocal Scoring forecasters discover in days how well they predicted other forecasters' 10-20 year predictions whereas objectively scored forecasters must wait 10-20 years for real-world accuracy feedback.

The Rapidity-of-Delivery curve heads in the opposite direction from the Expected Accuracy curve: it rises as a function of time to resolution because the more warning time a society has to implement promising policies, the better—at least up to a point. Setting lower bounds on rapidity of delivery is easy because no policy, no matter how sound, works instantaneously. An accurate prediction that we will be vaporized tomorrow by a massive solar storm will not be useful. Setting upper bounds is trickier. There is often fierce debate about which policies will work and how patient to be.[37] We assume here that a 50-year time horizon is sufficient for enacting most preemptive policies. And we model Rapidity as a logarithmic function of time-horizon: growing sharply as horizons extend from one to several years, and flattening beyond that point.

Figure 6 also predicts zones of peak policy utility for X-risk tournaments, zones we can estimate by multiplying the Expected Accuracy and Rapidity-of-Delivery functions. The left panel assumes faster decay of accuracy ($d = 0.75$); the right panel, slower decay ($d = 0.9$). So the left panel's zone of peak utility comes sooner than the right panel's: at 5 to 10 years versus 20-40 years. These assumptions reflect our view that it will prove possible to slow accuracy decay by using intersubjective methods, like Reciprocal Scoring, that give immediate accuracy feedback even for long-run outcomes—and thus encourage more rigorous data-driven forecasting strategies.

[37] Some policies, like banning gain-of-function virological research, may mitigate risks within months while others, like increasing funding for AI alignment or carbon capture research may have more delayed effects. Still, we expect the rapidity of delivery to matter less on the margin as time horizons expand beyond several decades. For example, discovering a 99.9% chance of a giant asteroid hitting Earth in 50 versus 100 years is unlikely to result in a much more effective policy response.

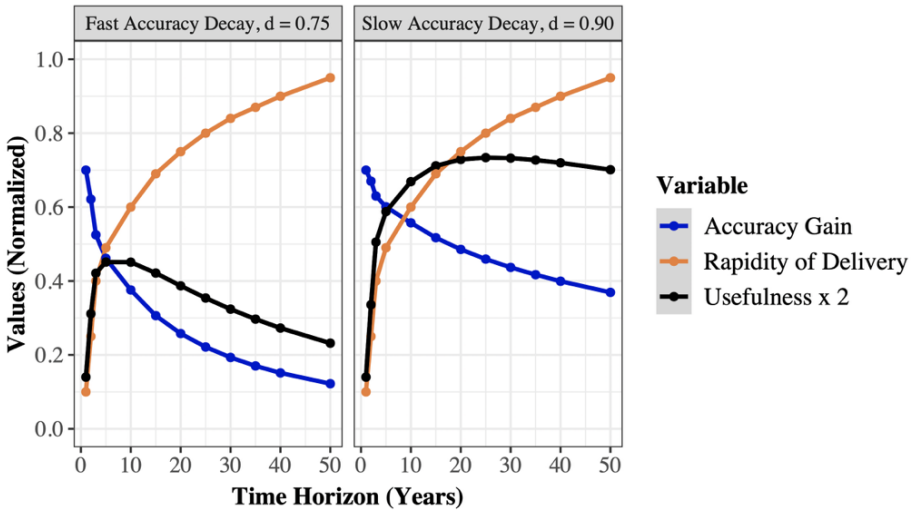

Figure 6: Hypothetical mapping of forecasting accuracy and rapidity of delivery of forecasts as a function of time. Projected usefulness, the product of accuracy and rapidity, is scaled up two times to enhance visibility. Left and right panels illustrate faster and slower accuracy decay rates, respectively, and how these relate to zone of maximal usefulness.

The sooner we launch X-risk tournaments, the sooner we can estimate actual, not just hypothetical, zones of peak usefulness. Facts can then inform our assumptions about key unknowns such as: (a) starting levels of accuracy; (b) the optimal weighting of forecasts incentivized by objective and intersubjective measures across time horizons; (c) the relationship between accuracy of optimal methods and time horizon.

Exact functional forms and parameter values matter: they have operational implications for coordinating the forecasting and risk-mitigation phases of X-Risk knowledge production. Our analysis in Figure 6 assumes that only one metric of good judgment, forecasting accuracy, is affected as time horizons stretch out. But we can reasonably posit that the three other gauges of good judgment in Figure 5 will also be affected: generating probative questions, persuasive rationales, and sensible policy priorities may all become harder as we examine more distant futures. Given that all three measures involve Reciprocal Scoring and thus immediate accountability feedback, let's posit these three measures will decay at the same rate as for intersubjective forecasts: 0.9. But this is, of course, an additional assumption in need of testing.

Figure 7 combines Figure 5's projections of Expected Information Gain from Reciprocal Scor-

ing with Figure 6's projections of the policy utility of X-risk tournaments as a function of Expected Accuracy of forecasts and Rapidity of Delivery. Imagine two policy-makers who live in a Figure 6 forecasting environment. One has the benefits of the Expected Information Gains in Figure 5; the other does not. Each must make resource-allocation decisions among X-risk priorities at multiple junctures over the next fifty years. At each juncture, only the better informed policy-maker has access to X-risk tournament data: objective accuracy metrics for recently resolved short-term questions as well as forecasts elicited with objective and intersubjective accuracy incentives across the entire 50-year horizon.

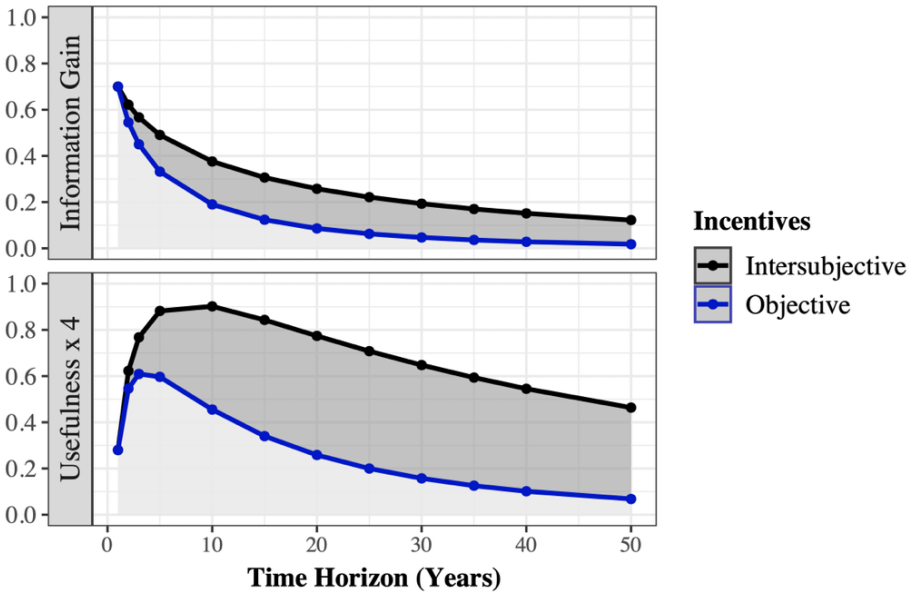

Figure 7: Hypothetical mapping of information gain and policy utility. Rapidity of delivery is held constant as in Figure 6. Projected usefulness, the product of accuracy and rapidity, is scaled up four times to enhance visibility. The light-gray area below the Objective line shows expected gains from running X-Risk tournaments with purely objective scoring incentives; the dark gray area shows the additional expected gains from introducing intersubjective incentives.

In the top panel of Figure 7, the y-axis measures information gain.[38] Zero represents the in-

[38]Ideally, we could define the y-axis of Figure 7 to answer the question that, we suspect, potential sponsors most want answered: How much do X-risk tournament data improve policy-makers' chances of investing in the right policy option, beyond the business-as-usual baseline? Or to use a framing popular among venture capitalists, how much can X-risk tournaments improve hit rates for spotting promising policies, holding false-positive rates constant? The answer is, unfortunately, profoundly context-specific. It hinges not only on the values and thinking styles of policy-makers but also on: (a) the options they deem politically feasible and on the table for adoption; (b) their perceptions of the net benefits of adoption; (c) the sensitivity of their estimates of net benefits to shifts in probabilistic forecasts. For all these reasons, we call the answer to this question the ultimate elusive metric.

formation available to the policy-maker in the business-as-usual world, with no access to X-risk tournament data. The light-gray area between zero and the objective-accuracy-only line captures cumulative information gain over time from access to forecasts from a basic first-generation tournament. The larger dark-gray area captures the additional cumulative information gain from access to forecasts using a combination of objective and intersubjective elicitation tools and incentives. The black line represents the expected information gain from running the multi-method X-Risk tournament proposed here.

The Bottom panel of Figure 7 combines estimates of Information Gain with Rapidity of Delivery (reusing values from Figure 6, not shown in Figure 7) to map the projected policy usefulness of forecasts that are only objectively scorable and of forecasts that include a mix of objectively scorable and intersubjectively scorable questions. The benefits of intersubjective metrics are larger in the Bottom Panel because we are multiplying the projected accuracy boost from including intersubjective metrics by the even larger projected timeliness boost from such metrics.

Our interpretation of Figure 7 is that policy-makers inhabiting a world with these functional properties are well advised to weight heavily: (a) short-run objectively resolvable 1-to-3-year forecasts on questions plucked from high-Evidence-Ratio conditional trees (Application 2 of Reciprocal Scoring) ; (b) the most persuasive arguments for those forecasts from the Hybrid Tournament (Application 3 of Reciprocal Scoring); (c) medium-run, 4-to-10-year intersubjectively scorable forecasts from the Shadowing-of-objective-forecasts exercises (Application #1 of Reciprocal Scoring); (d) estimates of most impactful policy options from the Risk-Mitigation Tournament (Application 4 of Reciprocal Scoring). They would also be well-advised to be extra cautious when time series of belief updates of short-to-medium-range forecasts exhibit excessive volatility and to be wary of very long-run forecasts, objective and intersubjective, because of the expected decay in accuracy with expanding time horizons.

All of this suggests that the longtermist community will need to come to existential terms with a mildly paradoxical possibility: the best way to promote its long-run goals is to focus on short-run forecasts from high-value conditional trees and on the most persuasive rationales for

those forecasts. Our collective survival may be better viewed not as one grand century-spanning challenge but as a succession of much shorter ones, in the 5 to-10 year range.

We suspect this message will be a hard sell. Regardless of their track records, charismatic long-view visionaries will continue captivating the popular imagination in ways that short-run incremental forecasters, toiling in the empirical trenches, can't match.[39] But we should recall that the "superforecasters" who prevailed in a decade of IARPA-monitored tournaments fit the latter, not the former, profile. If we were to venture another wager, it would be that they will also emerge as the "winners" of X-risk tournaments.

[39] In Expert Political Judgment, Tetlock (2005) noted a perverse inverse relationship between prominence and accuracy among the experts. The X-risk tournaments will test the robustness of this result on larger samples and on issues with more immediate policy stakes. The tournament will also test how fast regular forecasters learn to distinguish the predictive-value-added of explanations from a wide range of sources.

REFERENCES

Atanasov, P., Himmelstein, M. (2022) Talent Spotting in Crowd Prediction. Forthcoming in Seiffert, M. (Ed)., Judgment in Predictive Analytics, International Series in Operations Research Management Science Series, Springer NY.

Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., Ungar, L., & Mellers, B. (2017). Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Science, 63*(3), 691–706. https://doi.org/10.1287/mnsc.2015.2374

Atanasov, P., Witkowski, J., Mellers, B., & Tetlock, P. (2021). *Crowdsourced forecast elicitation: Methods vs. individuals* [video]. YouTube. https://www.youtube.com/watch?v=_DQAXWJVl2s

Atanasov, P., Witkowski, J., Ungar, L., Mellers, B., & Tetlock, P. (2020). Small steps to accuracy: Incremental belief updaters are better forecasters. *Organizational Behavior and Human Decision Processes, 160*, 19–35. https://doi.org/10.1016/j.obhdp.2020.02.001

Athey, S., Chetty, R., Imbens, G. W., & Kang, H. (2019). *The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely* (No. w26463). National Bureau of Economic Research. https://www.nber.org/papers/w26463

Augenblick, N., & Rabin, M. (2021). Belief movement, uncertainty reduction, and rational updating. *The Quarterly Journal of Economics, 136*(2), 933–985. https://doi.org/10.1093/qje/qjaa043

Berlin, I. (1950). *The hedgehog and the fox: An essay on Tolstoy's view of history*. Princeton University Press.

Bostrom, N. (2019). The vulnerable world hypothesis. *Global Policy, 10*, 455–477. https://doi.org/10.1111/1758-5899.12718

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. https://doi.org/10.1037/h0046016

Clemen, R., & Winkler, R. (1999). Combining probability distributions from experts in in risk analysis. *Risk Analysis, 19*, 187–203. https://doi.org/10.1111/j.1539-6924.1999.tb00399.x

Cvitanic, J., Prelec, D., Riley, B., & Tereick, B. (2019). Honesty via choice-matching. *American Economic Review: Insights, 1*(2), 179–192. https://doi.org/10.1257/aeri.20180227

Dana, J., Atanasov, P., Tetlock, P., & Mellers, B. (2019). Are markets more accurate than polls? The surprising informational value of "just asking". *Judgment and Decision Making, 14*(2), 135–147.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*(7), 571. https://doi.org/10.1037/0003-066X.34.7.571

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, *11*(4), 227–268. https://doi.org/10.1207/S15327965PLI1104_01

Frank, M. R., Cebrian, M., Pickard, G., & Rahwan, I. (2017). Validating Bayesian truth serum in large-scale online human experiments. *Plos One, 12*(5), e0177385. https://doi.org/10.1371/journal.pone.0177385

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*(2), 155–170.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist, 52*(1), 45–56. https://doi.org/10.1037/0003-066X.52.1.45

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*(1), 1–38. https://doi.org/10.1207/s15516709cog0702_3

Hammond, K. R., & Adelman, L. (1976). Science, values, and human judgment: Integration of facts and values requires the scientific study of human judgment. *Science, 194*(4263), 389–396. https://doi.org/10.1126/science.194.4263.389

Harvey, C., Rattray, S. & van Hemert, O. (2021). *Strategic risk management: Designing portfolios and managing risk*. New York; Wiley.

Himmelstein, M., Budescu, D., Han, Y. (2021). The Wisdom of Timely Crowds. Working Paper.

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences, 101*(46), 16385–16389. https://doi.org/10.1073/pnas.0403723101

Jervis, R. (1976). *Perception and misperception in international politics*. Princeton University Press.

Kahneman, D. (2011). *Thinking, fast and slow.* Macmillan.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263–292. https://doi.org/10.1126/science.194.4263.389

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: a flaw in human judgment*. Little, Brown Spark publishing.

Karger, E., Monrad, J., Mellers, B., & Tetlock, P. (2021). *Reciprocal scoring: A method for forecasting unanswerable questions*. Working Paper.

Kotovsky, K., Hayes, J. R., & Simon, H. A. (1985). Why are some problems hard? Evidence from Tower of Hanoi. *Cognitive Psychology, 17*(2), 248–294. https://doi.org/10.1016/0010-0285(85)90009-X

Landeta, J. (2006). Current validity of the Delphi method in social sciences. *Technological Forecasting and Social Change*, *73*(5), 467–482. https://doi.org/10.1016/j.techfore.2005.09.002

Lepper, M. R., Greene, D., & Nisbett, R. E. (1973). Undermining children's intrinsic interest with extrinsic reward: A test of the "overjustification" hypothesis. *Journal of Personality and Social Psychology*, *28*(1), 129–137. https://doi.org/10.1037/h0035519

Liu, Y., Wang, J., & Chen, Y. (2020). *Surrogate scoring rules* [Paper presentation]. Proceedings of the 21st ACM Conference on Economics and Computation, Virtual Event, Hungary.

McCoy, J., & Prelec, D. (2017). A statistical model for aggregating judgments by incorporating peer predictions. *arXiv preprint,* arXiv:1703.04778.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* University of Minnesota Press.

Millner, A., & Heyen, D. (2021). Prediction: The long and the short of it. *American Economic Journal: Microeconomics*, *13*(1), 374–398. https://doi.org/10.1257/mic.20180240

Murphy, A & Winkler, R. (1984) Probability Forecasting in Meteorology, *Journal of the American Statistical Association, 79*(387), 489–500. https://doi.org/10.1080/01621459.1984.10478075

Myatt, D. P., & Wallace, C. (2012). Endogenous information acquisition in coordination games. *Review of Economic Studies, 79*(1), 340–374. https://doi.org/10.1093/restud/rdr018

Nagel, T. (1989). *The view from nowhere*. Oxford university press.

Ord, T. (2020). *The precipice: Existential risk and the future of humanity.* Hachette.

Page, S. (2006). *The difference. How the power of diversity creates better groups, firms*. Princeton University Press.

Patil. S., & Tetlock, P.E. (2014). Punctuated incongruity: A new approach to managing trade-offs between conformity and deviation. In B. Staw, & A. Brief (Eds.), *Research in organizational behavior* (pp. 155–171). JAI Press.

Prelec, D. (2004). A Bayesian truth serum for subjective data. *Science, 306*(5695), 462–466. https://doi.org/10.1126/science.1102081

Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature, 541*(7638), 532–535. https://doi.org/10.1038/nature21054

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Sagan, S., & Waltz, K. (1995). *The spread of nuclear weapons: A debate.* Norton

Satopää, V. A., Salikhov, M., Tetlock, P. E., & Mellers, B. (2021). Bias, information, noise: The BIN model of forecasting. *Management Science,* Advance online publication.

Simon, H. A., & Hayes, J. R. (1976). The understanding process: Problem isomorphs. *Cognitive Psychology, 8*(2), 165–190. https://doi.org/10.1016/0010-0285(76)90022-0

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, New Series, 103*(2684), 677–680. https://doi.org/10.1126/science.103.2684.677

Taleb, N. N., & Tetlock, P. E. (2013). *On the difference between binary prediction and true exposure: With implications for prediction markets and forecasting tournaments*. Berkeley Statistics. https://www.stat.berkeley.edu/~aldous/157/Papers/taleb_tetlock.pdf

Tarsney, C. (2019). *The epistemic challenge to longtermism*. philpapers. https://philpapers.org/archive/TARTEC-2.pdf

Tetlock, P. (2005). *Expert political judgment: How good is it? How can we know*? Princeton University Press.

Tetlock, P. E. (2017). *Expert political judgment: How good is it? How can we know?* (New ed.). Princeton University Press.

Tetlock, P. E., & Belkin, A. (1996). *Counterfactual thought experiments in world politics: Logical, methodological, and psychological perspectives*. Princeton University Press.

Tetlock, P. E., & Gardner, D. (2015). *Superforecasting: The art and science of prediction*. Random House.

Tversky, A., & Kahneman, D. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 363–391. https://doi.org/10.1038/nature21054

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323. https://doi.org/10.1007/BF00122574

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101*(4), e547. https://doi.org/10.1037/0033-295X.101.4.547

Waggoner, B., & Chen, Y. (2014, September). *Output agreement mechanisms and common knowledge* [Paper presentation]. Second AAAI Conference on Human Computation and Crowdsourcing, Pittsburgh, Penn.

Witkowski, J., Atanasov, P., Ungar, L., & Krause, A. (2017). *Proper proxy scoring rules* [Paper presentation]. Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA.

TECHNICAL APPENDIX

Drawing directly from the theoretical justification and text of Karger et al. (2020), suppose we ask three forecasters—Anna, Bob, and Charlie—a question unresolvable in their lifetimes: what will be the death toll from pandemics in the next century? Will they do the right cognitive thing or invent sneaky, low-effort ways to game the system? Each forecaster must submit forecasts without knowing the others' strategies. Let's limit strategies to three: each can submit a low-effort forecast of 0, a medium-effort forecast of 100 million (say, the result of a quick Google search), or a high-effort forecast of 500 million after serious modeling. If the forecasters choose the same forecast, they get the accuracy prize; otherwise 0. That now leaves three simple, pure-strategy equilibria:

(1) Anna, Bob, and Charlie all submit the low-effort forecast of 0

(2) Anna, Bob, and Charlie all submit the medium-effort forecast of 100 million

(3) Anna, Bob, and Charlie all submit the high-effort forecast of 500 million

Researchers do not know the high-effort forecast, 500 million, is correct: that is the answer they seek. And in all three equilibria, no forecaster should stray because doing so reduces her prize from a positive value to zero.[40] Reciprocal Scoring needs revising. It has two sub-par equilibria.

Fortunately, we can foreclose the undesirable equilibria. We "just" need to persuade forecasters that: (a) at least one among them—say, Anna—is a good-faith forecaster who will invest high effort; (b) they will converge on the same answer if they do high-effort searches (the "common prior" assumption). Now only one pure equilibrium remains: the high-effort one.[41] Any other forecasting strategy yields a lower expected prize.[42]

[40] Assume the prize pool is between \$0 and \$100 for each forecaster and is not a function of the prizes that other forecasters receive, which would complicate things.

[41] But are there mixed strategy equilibria? No: if we assume there is at least one forecaster, Anna, who adopts the high-effort forecasting strategy in all states of the world. In that case, Bob and Charlie cannot improve their monetary prize by adopting a mixed (random guessing) strategy because it would yield a monetary reward of zero in some states of the world and leave them better off forecasting the high-effort equilibrium in all states of the world.

[42] This basic model generalizes to $N > 3$ forecasters. Assuming that there is an "Anna" who will always submit a high-effort forecast, the only equilibrium is for all other forecasters to do so as well, regardless of whether there are 10 or 10 million forecasters. This is true even if non-Anna forecasters do not know the identity of the high-

Reaching that one pure equilibrium state in an actual experiment, not just a thought experiment, is however nontrivial. We must anticipate the real-world temptations to deviate from the truth-seeking equilibrium—and pre-empt them. We prioritize four precautions:

(1) Forecasters who expect others to be biased might not submit true-belief forecasts. Reciprocal Scoring depends on each team assuming the other is pursuing epistemic goals and is unbiased by stereotypes of the other team's priors. The logic breaks down if we don't block the shortcut of substituting hard work with glib guesses (e.g., "those crazy China hawks are obsessed about a bio-war so let's just mirror their craziness"). To block this tactic, each team must believe the other has a strong track record and cares, above all, about preserving that reputation. Our solution here is to rely on superforecasters, a group of forecasters who earned status in past tournaments by making surprisingly accurate judgments on a wide array of questions.

(2) Forecasters with private information might not submit true-belief forecasts. If a forecaster knew of a brilliant article in an obscure language, she may opt to ignore it. Our solution is to deploy teams that have free internal flows of information and are large enough that the likelihood of private useful information on only one team drops fast. Suppose the forecaster puts a 10% chance on another discovering the article. In a 100-person tournament with two 50-person teams, this yields only a $0.9050 = 0.5\%$ chance that the other team would miss the article, which gives the forecaster reason to factor it into her forecast.

(3) Forecasters might be lazy. Suboptimal effort is hard to police so our solution has three parts: (a) recruiting only forecasters who value their reputations for integrity and rigor; (b) requiring regular visible contributions to team debates; (c) hosting within-team competitions to generate rationales that judges evaluate on creativity and accuracy. Given the can-do

effort forecaster. We tweak this setup operationally. We divide forecasters into two independent teams and give each forecaster a prize inversely proportional to the squared distance between individual forecasts and the median forecast on the other team. We also rely on teams that value open dissent, instead of individual forecasters, which should also boost the quality of forecasts (Tetlock & Gardner, 2015).

efforts of top teams in past tournaments, we are optimistic about nudging teams toward progressively higher-effort equilibria.

(4) Forecasters might be dishonest. If a forecaster befriended someone on the other team, the two could coordinate to maximize their prizes with minimal effort. Our solution is, again, threefold: (a) tightening anonymity of team membership; (b) monitoring deliberations to detect bias or sloppiness; (c) expanding the set of teams beyond two and evaluating forecasters on the distance between their forecasts and other teams' aggregate forecasts — or the distance between their forecast and randomly selected individuals on other teams.

Karger et al. (2021) propose a formal theoretical model of Reciprocal Scoring that justifies these ground rules—rules that people might otherwise see as arbitrary.[43]

To formalize the theoretical basis for Reciprocal Scoring, consider the following model, drawing from the framework of Myatt & Wallace (2012). A tournament organizer wants to know the best possible forecast (an unknown real number θ) about an event. θ could represent a forecast of coronavirus deaths in the world or the global population in 2030. A continuum of forecasters perform the following steps:

1. Each forecaster pays a cost C(e) to exert effort 'e' and obtains an independent draw of information θ_i about the world from a normal distribution with mean θ and variance $\frac{1}{e}$.[44] The signal is unbiased because its distribution is centered on the true value θ. So, the more effort forecasters exert, the more likely it is their draw from this information source will yield a signal closer to the truth. The cost function C(e) is assumed to be increasing, convex, and differentiable. Intuitively, the cost function could represent cost from the time spent gathering information.

2. The forecaster submits a forecast 'f' to the tournament organizer.

[43] Karger et al. (2021) see their work as building on that of Myatt & Wallace (2012).

[44] Myatt & Wallace, 2012, we assume that θ has an improper prior, with the same prior probability given to any value of θ.

3. The forecaster receives unobserved utility $U = -\alpha * (f-\theta)^2 - \pi * \left(f-\overline{f}\right)^2 - C(e)$. In this utility function, $\alpha > 0$, is a fixed (unobserved) parameter that reflects the intrinsic importance of the truth to each forecaster, and $\pi > 0$ is a prize that the tournament organizer pays to participants as a function of the difference between their forecast and the average forecast of other forecasters; $\overline{f}$ is the average of all forecasters' forecasts, as received by the tournament organizer, in the style of Reciprocal Scoring.[45]

In this model, a forecaster's best option after selecting effort 'e', in response to any signal, θ, is to submit a forecast 'f' that maximizes her own utility, which is done by choosing 'f' such that the first derivative of the utility function with respect to f is zero–in other words, $\alpha * (f - E[\theta \mid \theta_i]) + \pi * \left(f - E\left[\overline{f} \mid \theta_i\right]\right) = 0$. The forecaster accomplishes this by submitting a forecast f equal to $\frac{\alpha}{\alpha+\pi} * E[\theta \mid \theta_i] + \frac{\pi}{\alpha+\pi} * E\left[\overline{f} \mid \theta_i\right]$. In other words, the forecaster will forecast a weighted average of her best estimate of the value of θ given her personal signal θ and her expectation of what the other forecasters will forecast (again, given her personal signal θ). Because each forecaster only receives one unbiased signal about the true value of the world (θ) and the other forecasters do as well, this weighted average corresponds to submitting a forecast of her own signal, θ.

In this signaling model, a forecaster cares at least marginally about two quantities: how close her forecast is to the truth (intrinsic motivation) and how close her forecast is to the average forecasts of other forecasters–a quantity we incentivize directly using Reciprocal Scoring and a prize π. This incentivizes the forecaster to submit her best estimate of the truth.

To go a step further–how much effort will a forecaster exert? She will exert effort to maximize her own utility, which involves internalizing a tradeoff between the increased cost of obtaining a precise signal (channeled through C(e)) and the increased payoff of submitting a forecast that is closer to both the unobserved true value (θ) and to the average forecast submitted by other forecasters. In the absence of Reciprocal Scoring, where forecasters care only about the distance between their forecast and θ, tournament organizers have no way of incentivizing effort because

[45] In our empirical analyses of Reciprocal Scoring we use the median and not the average of other forecasters' forecasts to avoid a case where a small number of outliers affects our results.

θ is unobserved. In this model, where Reciprocal Scoring creates an incentive for forecasters to report forecasts similar to other unbiased forecasters, the tournament organizer can increase the accuracy of forecasts by reducing the cost of effort or by increasing the payoff from exerting effort (which is equivalent to increasing π). Practically, this could involve offering large prizes for forecaster effort. Because each forecaster's signal is independent conditional on the true value of theta, the median or average of all forecasters' forecasts is more precise than any individual forecast, and the precision of the median or average of all forecasts is increasing in e (and therefore increasing in the value of the prize). So, a tournament organizer with large incentives to maximize accuracy will use prizes to encourage a high level of individual effort from forecasters.