Welcome to the Hybrid Forecasting-Persuasion Tournament

Dear Colleagues,

This tournament will present long-run and shorter-run questions bearing on existential risks. You will forecast in three phases: (1) independently; (2) as part of a team comprised of [Superforecasters / subject matter experts]; and (3) as part of a team comprised of both Superforecasters and subject matter experts.

Throughout, feel free to use any references and consult anyone, with two exceptions: (1) while forecasting independently, do *not* consult other participants, even if they are on your team; (2) while forecasting as a member of a team, do *not* consult members of other teams.

Separating participants in the Independent-Work Phase is essential to our research design. Your independent forecasts at the start will establish a baseline that lets us assess "process gain" attributable to each later phase of information-sharing and collaboration. And your updated forecasts will reveal how much you have benefited from seeing others' forecasts and rationales. Please help us isolate which stages of the experiment add the most value.

For the same reasons, please refrain during the tournament from blog posts and other messaging that leaks information about your forecasts and rationales, or that of teammates, to participants in other conditions. By participating, you are agreeing to postpone such activities until we complete all tournament activities—roughly four months after the conclusion of the tournament.

It is crucial that everyone feel comfortable expressing their true beliefs. You have the option of forecasting anonymously—and by signing on you are agreeing not to name teammates or discuss their comments with non-participants without your teammates' explicit permission. The research team will anonymize all forecasts and rationales for data-analysis and ensure you cannot be linked to your forecasts and rationales.

All work will be asynchronous via our online platform so you can forecast whenever convenient. Note though that the stages of independent and team forecasting will be time-limited so everyone will need to move from one stage to the next according to the tournament schedule. You may update your forecasts and rationales as often as you like in each stage. **Your *final* forecast in each stage will be the one that we score.**

You will encounter two categories of probabilistic forecasting questions. Category-1 forecasts are objectively resolvable questions about early-warning indicators. We will ask about distributional beliefs (5^{th} 25^{th}, 50^{th}, 75^{th}, and 95^{th} percentiles) of quantities such as, say, the number of nuclear powers in 2024. Category-2 forecasts are about long-run risks to society. Here we will use a new "intersubjective" metric of accuracy, Reciprocal Scoring, that challenges

forecasters to predict the forecasts of a panel of "Superforecasters" who, in principle, care solely about accuracy.[1] We discuss these rules below and in a stand-alone scoring document.[2]

Some Category-2 forecasts are probabilistic (Category-2a), with values that could range from the microscopic (0% to 0.1%) to the ominously substantial (10%, 20% or higher). Others involve continuous quantities (Category-2b)—such as the 5th, 25th, 50th, 75th, and 95th percentile forecasts of the year when humanity will go extinct. Due to the long time horizons, we cannot objectively verify Type 2 forecasts—hence the need for the intersubjective tool of Reciprocal Scoring.

Stage 0 - Onboarding (early May):

We will reach out to interested participants 10-14 days before the tournament begins. You will be given a short survey and an informed consent on Qualtrics. You will also receive an email that explains how to access the forecasting platform and create an account. **Please contact us immediately if you have not yet received that information as the tournament start date approaches or are having problems with the platform.**

Your intake survey will elicit your baseline beliefs about existential and global catastrophic risk. At the end of the tournament, we will ask you these questions again to see if and how your beliefs changed.

We recognize the format of this tournament will be unfamiliar. To ease your way, we will hold two onboarding webinars (one for all Superforecasters; the other for all domain experts) so that everyone has a chance to talk with the researchers about the tournament. Each onboarding webinar will be recorded for those who cannot join the live session.

Before the webinar, please log onto the platform. We've loaded a few practice questions so you can learn about the interface. Accessing the platform before the webinar will be helpful because our software team will be on hand to answer platform-related questions.

Of course, you'll have access to the in-platform help feature throughout the onboarding process and the tournament itself. We hope to respond to most queries within 1 business day but can't guarantee responses outside regular business hours. We will provide extra help-desk personnel at the start of the tournament and during the transition points between stages to minimize the wait-time for assistance when it is most likely to be needed.

We also want to provide everyone with a common foundation on principles of probabilistic forecasting and tips for team collaboration. You should already have received an email on how

[1] For more information, see Karger, Ezra, Joshua Monrad, Barb Mellers, and Philip Tetlock. "Reciprocal Scoring: A Method for Forecasting Unanswerable Questions." *Available at SSRN* (2021).

[2] Both this operational plan and the stand-alone scoring document will be available via links on the tournament forecasting platform so that you can always readily review the information.

to access those training materials. There is also a link to training materials on your platform so that you can refer back to them at any time.

The platform also contains information on how to report potential information hazards to tournament organizers. An infohazard is information whose dissemination could cause harm.[3] We have developed an infohazard-avoidance plan after consulting with experts at Open Philanthropy and other organizations that research existential risks. If you see information that might be infohazardous, please tell us so we take appropriate action.

Stage 1 – Initial Forecasts Working Alone. Crucial Baseline Data (three weeks):

You will have one week to make independent forecasts on around ten (10) long-run questions and post a rationale for each forecast and sources on which you are relying. Again, please do not consult other participants but feel free to consult anyone outside the tournament.

You will then have another two weeks to forecast on shorter-run questions, roughly 50 of them. We do NOT expect anyone to be able to offer a thoughtful forecast and rationale on every question. We do though need a baseline assessment of the opinions you hold at the start, even if just guesses.

To keep workload for these baseline assessments manageable, we recommend "triage" shortcuts. As you read through the questions, sort them into three categories: (a) questions about which you know a lot and can estimate an answer quickly, without further research; (b) those about which you feel somewhat knowledgeable and could estimate an answer with a quick internet search or asking trusted colleagues outside this tournament; (c) those about which you feel quite clueless and are very reluctant to venture a forecast. For this last category of questions, feel free to check the "just-guessing" box.

Establishing a baseline of starting opinions will enrich the scientific database as well as let us give you detailed feedback on how your (anonymized) profile of judgments evolved in response to the debates inside the tournament.

After you submit your forecasts on these shorter-run questions, you will be able to see your teammates' forecasts and rationales. Feel free to update your own beliefs at this point. Be aware that we expect that initial forecasts will continue to trickle in—so stage 2 offers a fuller opportunity for conversations and updating.

Note: We need to know your reasoning, so we will offer ten $2,000 prizes for the most persuasive rationales across all questions. We will evaluate persuasiveness after the tournament by showing the highest-quality rationales to numerate non-participants who will submit

[3] For more information, see Bostrom, Nick. "Information hazards: a typology of potential harms from knowledge." *Review of Contemporary Philosophy* 10 (2011): 44-79.

forecasts before and after seeing the rationale. We will give prizes to those rationales that help non-participants to update their forecasts in the direction of greater accuracy.

Stage 2 – Team-Based Forecasting Begins (three weeks):

You have now joined your team of fellow [Superforecasters/SMEs]. You can see each other's Stage-1 forecasts and rationales on all questions—and can decide whether to modify your opinions. To minimize information overload, we recommend focusing on only a few questions at a time.

You can also do more than just look at teammates' judgments. You can talk with them: question, debate and collaborate. We do not expect teams to agree on a single "consensus" forecast. You can still post your own forecasts and rationales, taking account of others' views as you see fit. We do though encourage teams to use classic techniques for improving deliberations: precision questioning, devil's advocacy/red-teaming, and adversarial collaboration (described in optional training documents).

Even though you are working asynchronously, we encourage interaction and will reward you for (1) replying to other's forecasts and rationales and (2) updating your own forecasts and rationales.

Stage 3 - Interaction within Teams That Combine Superforecasters and Domain Experts (three weeks):

In Stage 3, we give Superforecasters and subject matter experts their first opportunity to collaborate. Your Stage 2 [Superforecaster / subject matter expert] team will join forces with one of the Stage 2 [subject matter expert / Superforecaster] teams to form a new team with roughly twice as many people. You can now see all of the forecasts and rationales previously posted by the Superforecasters and subject-matter experts who are members of your new, expanded team. Please collaborate with your combined team and modify your forecasts and rationales as you see fit.

Stage 4 – Creating Team Rationales (two weeks):

Operating within the combined Superforecaster/subject-matter-expert teams formed in Stage 3, you will craft wiki-style rationales for each question (specializing on what you know best). These rationales will reflect the "team" position; however, again, we do not require consensus and you are encouraged to develop minority viewpoints.

Tournament organizers will designate one team member to serve as the "editor" for each question's rationale. On average, each team member will be assigned as editor for two questions total. All team members will be able to see that rationale as it is drafted and are encouraged to contribute. For each question, we will compensate both the editor and team members who make major contributions.

Wiki editors for a team will be compensated for producing wikis that persuade a second team of Superforecasters and subject-matter-experts to adjust their forecast toward the truth (for short-run questions) and toward the first team's forecasts (for long-run questions).

Stage 5 – Testing the Persuasive Power of Team Rationales (two weeks):

Stage 5 tests the persuasive power of the wiki-style rationales each team drafted in Stage 4. You will remain in your Stage-4 teams but can now see another team's rationales for all questions. You can update your forecasts and rationales based on all the information available.

Stage 6 – Exit Survey, and Updating Your Allocations to Risk-Prevention/Mitigation Efforts

We will ask you to retake the intake survey so we can assess how your views have evolved. You also will have an opportunity to update your proposed allocation of funds to various risk-prevention or mitigation efforts.

Benefits of Participation in This Tournament

Our hope is that participation will give you a unique opportunity to think deeply about many of the biggest challenges facing humanity over the next century. Tournament organizers have worked with experts to target key uncertainties. While this is an experiment, we see each team as a working group—and we fully expect your final reports will be read with interest by influential actors in the Effective Altruism community and beyond.

Your teams will consist of both subject-matter-experts on existential risk and talented generalists—Superforecasters—with proven track records of making accurate probabilistic forecasts. Through this experience, you may expand your professional networks and discover new collaboration opportunities. At the very least, you can expect fun conversations with a diverse group of thoughtful people.

You are also advancing the science of forecasting. The ideas we are testing here build on decades of research on how to boost human forecasting accuracy. We view every forecaster as a collaborator and potential co-author on publications emerging from the tournament.

The rules in this tournament are not arbitrary; they enable testing ideas that would otherwise be untestable. Thank you for helping us push the science forward.

The more you engage, the more you will learn about techniques for improving your effectiveness in other environments. To reward active engagement, each team will vote to recognize teammates who were:

1. most helpful overall in facilitating the team's deliberations and report generation;

2. most effective in critiquing and responding to other teammates' contributions (adversarial collaborations); and
3. most persuasive in presenting rationales that helped the team arrive at a better understanding of key topics.

The three teammates receiving the highest votes in each of these categories will be invited to attend a virtual conference with the tournament organizers, led by Phil Tetlock, to discuss the findings and directions for future research.

Of course, we will also compensate each participant for effort and forecasting accuracy.

Effort includes:

- Posting an initial probabilistic forecast and accompanying rationale for each of the ten long-run questions;
- Posting an initial probabilistic forecast and accompanying rationale for as few or as many of the short-run questions as each participant chooses;
- Updating forecasts and rationales for all questions in response to:
    - seeing forecasts and rationales from other members of one's Stage 1 [Superforecaster / subject-matter-expert] team;
    - seeing forecasts and rationales from other members of one's Stage 2 combined Superforecaster / subject-matter-expert team; and
    - seeing the wiki-style rationales from a different Superforecaster / subject-matter-expert team;
- Editing the wiki-style "team" rationale for two forecasting questions;
- Making a significant contribution to the wiki-style rationale for as few or as many of the remaining questions as each participant chooses; and
- Writing the most persuasive rationales in the first (individual) stage of the tournament.

Accuracy reflects:

- The final *individual* forecast for each forecast standing at the end of each stage; and
- The final *team* forecast for each forecast standing at the end of each stage from Stage 2b onward. For purposes of determining team accuracy, we will construct a team forecast for each question as the simple unweighted-median of the last standing forecasts of each team member who has made a forecast on the question.

We view participants in this tournament as collaborators and want you to feel fairly compensated for all your work. We plan to pay every participant who participates in this tournament at least $2,000 and, in cases of exceptional contributions, up to $10,000. We will circulate additional details about compensation closer to the start of the tournament. Participants will need to spend at least 40 hours on the tasks described above, over roughly three months. But we understand that participation will vary at an individual level and plan to provide the most compensation to those who participate the most.

Scoring

We pose two types of probabilistic forecasts. *Category-1 forecasts* are objectively resolvable, shorter-term questions about early warning indicators where we plan to elicit forecasters' distributional beliefs (5^{th} 25^{th}, 50^{th}, 75^{th}, and 95^{th} percentile forecasts) of quantities like the number of countries with a nuclear warhead in the year 2024. *Category-2 forecasts* are about long-run risks to society—and here we will measure accuracy using a new intersubjective metric that we call "Reciprocal Scoring" that challenges forecasters to predict the forecasts of a panel of "Superforecasters" not on your team who theoretically care solely about accuracy.

When Category-2 forecasts are probabilistic (Category 2a), we expect these probabilities to range from the microscopic (.0% to 0.1%) to the ominously substantial (10%, 20% or higher). When Category-2 forecasts involve continuous quantities (Category 2b)—like the 5^{th}, 25^{th}, 50^{th}, 75^{th}, and 95^{th} percentile forecasts of the year when humanity will go extinct. Because of the long time horizons, we will not have access to ground-truth verification of Category-2 forecasts.

Category 1: Distributional forecasts on short-term resolvable questions

These forecast estimates are elicited as distributions, e.g., the number of deaths from cause A, such that the forecaster is X% confident the number will not exceed a value of Y. We deal with forecast distributions later on. We start with the simplest case, in which we have derived from each forecaster a forecast (Y) of the number of deaths from cause A such that the forecaster is 50% confident the number of deaths will not exceed Y.

In this case, we propose a simple quadratic scoring rule—we will evaluate forecasters using the squared distance between their forecast and the true value, i.e., mean squared error.

We now move on to (briefly) describe the case where we elicit additional distributional forecasts. Our current plan is to elicit distributions at several (up to 5) quantiles. Forecasts for each quantile will be evaluated using the S-score measure (Jose & Winkler, 2022).

$$S\left(q_{\alpha}, x\right) = \alpha\, max\left(x - q_{\alpha}, 0\right) + (1 - \alpha)\, max(q_{\alpha} - x,\ 0).$$

The S-score is a function of the forecasted quantity (*q*) for quantile α, (q_{α}), and the realized value *x*. Lower scores are better, and the best score for a quantile is zero.

Assume we elicit forecasts of these percentiles (5%, 25%, 50%, 75%, 95%) for a given quantity (number of deaths due to some cause by 2024), and forecaster A gives us forecasts of (1, 5, 20, 200, 250) and forecaster B gives us forecasts of (20, 25, 50, 200, 250). Let us assume the true value is 20. The forecaster will receive an S-score of 0 for their median forecast, and positive scores for other quantile forecasts. Note that this scoring function is asymmetric for non-median forecasts: errors in which the realized values are more extreme than predicted (e.g., below the 10^{th} or above the 90^{th} quantile forecast values) incur higher penalties.

Category-2: Probabilistic and distributional forecasts verified using reciprocal scoring
Now consider unresolvable questions about long-run outcomes.

First (Category 2a), consider forecasts of long-run probabilistic questions (e.g., the probability that deaths from cause X will exceed 1,000).

We define the two variants of reciprocal scoring (RS) as follows. Let f_a be the probability forecast value for forecaster a, and f_s be the Superforecaster consensus, that is, the aggregate forecast derived from Superforecaster individual estimates.

RS Brier Rule = $2 \times [f_s - f_a]^2$, ranging from 0 (best) to 2 (worst)
RS Logarithmic Rule = $\ln (1 - |f_s - f_a|)$, ranging from -∞ (worst) to 0 (best)

The RS Brier Rule takes the value of 0 when the Superforecaster consensus is identical to the individual estimate; the maximum value is only reachable if the consensus probability is 0% and the individual forecaster predicts 100%, or vice versa. Similarly, according to the RC Logarithmic rule, an individual forecaster earns the best possible score (zero) if their estimate is identical to the Superforecaster consensus. If the consensus is 0%, and the individual estimates 100% (or vice versa), she would earn the worst possible score of -∞. As a practical matter, consensus estimates are unlikely to reach exactly 0%.

Second (Category 2b), consider forecasts of distributional quantities—like the forecaster's 5th, 25th, 50th, 75th, and 95th percentile forecasts of the number of deaths due to nuclear war by 2100.

Let F_{aq} be the forecast value for forecaster a at quantile q and let F_{sq} be the Superforecaster consensus at quantile q, that is, the median forecast derived from Superforecaster individual estimates. We propose a quadratic rule comparable to the Brier rule above:

RS Quadratic Rule = $\frac{1}{5} \sum_{q=0}^{5} [F_{sq} - F_{aq}]^2$

RS Quadratic Rule score is calculated as the squared deviation between an individual estimate and Superforecaster consensus, averaged across quintiles.