Forecasting Research Institute

Conditional Trees:

A Method for Generating Informative Questions about Complex Topics

AI Risk Case Study

Authors: Tegan McCaslin, Josh Rosenberg, Ezra Karger, Avital Morris, Molly Hickman, Otto Kuusela, Sam Glover, Zach Jacobs, Phil Tetlock

First released on: August 12, 2024
FRI Working Paper #3

Conditional Trees: A Method for Generating Informative Questions about Complex Topics[1]

Tegan McCaslin,* Josh Rosenberg,* Ezra Karger,*† Avital Morris,* Molly Hickman,* Otto Kuusela,* Sam Glover,* Zach Jacobs,* Phil Tetlock*‡

Abstract

We test a new process for generating high-value forecasting questions: asking experts to produce "conditional trees," simplified Bayesian networks of quantifiably informative forecasting questions. We test this technique in the context of the current debate about risks from AI. We conduct structured interviews with 21 AI domain experts and 3 highly skilled generalist forecasters ("superforecasters") to generate 75 forecasting questions that would cause participants to significantly update their views about AI risk. We elicit the "Value of Information" (VOI) each question provides for a far-future outcome—whether AI will cause human extinction by 2100—by collecting conditional forecasts from superforecasters (n=8).[2] In a comparison with the highest-engagement AI questions on two forecasting platforms, the average conditional trees-generated question resolving in 2030 was nine times more informative than the comparison AI-related platform questions (p = .025). This report provides initial evidence that structured interviews of experts focused on generating informative cruxes can produce higher-VOI questions than status quo methods.

[1] This research would not have been possible without the generous support of Open Philanthropy. We thank the research participants for their invaluable contributions. We greatly appreciate the assistance of Page Hedley, Kayla Gamin, Leonard Barrett, Coralie Consigny, Adam Kuzee, Arunim Agrawal, Bridget Williams, and Taylor Smith in compiling this report. Additionally, we thank Benjamin Tereick, Javier Prieto, Dan Schwarz, and Deger Turan for their insightful comments and research suggestions.

* Forecasting Research Institute

† Federal Reserve Bank of Chicago

‡ Wharton School of the University of Pennsylvania

[2] We will refer to this set of forecasters as "superforecasters" henceforth. Note that while seven of the forecasters are Superforecasters™ as officially designated by Good Judgment Inc., one is a skilled forecaster who does not have that label but has a comparable track record of calibrated forecasts.

Table of Contents

Executive summary

Introduction

From May 2022 to October 2023, the Forecasting Research Institute (FRI) (a)[3] experimented with a new method of question generation ("conditional trees"). While the questions elicited in this case study focus on potential risks from advanced AI, the processes we present can be used to generate valuable questions across fields where forecasting can help decision-makers navigate complex, long-term uncertainties.

Methods

Researchers interviewed 24 participants, including 21 AI and existential risk experts and three highly skilled generalist forecasters ("superforecasters"). We first asked participants to provide their personal forecast of the probability of AI-related extinction by 2100 (the "ultimate question" for this exercise).[4] We then asked participants to identify plausible[5] indicator events that would significantly shift their estimates of the probability of the ultimate question.

Following the interviews, we converted these indicators into 75 objectively resolvable forecasting questions. We asked superforecasters (n=8) to provide forecasts on each of these 75 questions (the "AICT" questions), and forecasts on how their beliefs about AI risk would update if each of these questions resolved positively or negatively. We quantitatively ranked the resulting indicators by Value of Information (VOI), a measure of how much each indicator caused superforecasters to update their beliefs about long-run AI risk.

To evaluate the informativeness of the conditional trees method relative to widely discussed indicators, we assess a subset of these questions using a standardized version of VOI, comparing them to popular AI questions on existing forecasting platforms (the "status quo" questions). The status quo questions were selected from two popular forecasting platforms by identifying the highest-engagement AI questions (by number of unique forecasters). We present the results of this comparison in order to provide a case study of a beginning-to-end process for producing quantitatively informative indicators about complex topics. (More on methods)

[3] To ensure the integrity of links in this report, we include stable archive.org links in parentheses after each citation to an external URL.
[4] More specifically, the ultimate question was defined as the global human population falling below 5,000 individuals at any time before 2100, with AI being a proximate cause of such reduction.
[5] "Plausible" meaning that the forecaster deemed the indicator event to be at least 10% likely to occur. This 10% probability was not necessarily an unconditional probability, but may have been conditional on a previous node in the conditional tree.

Results

The conditional trees method can generate forecasting questions that are more informative than existing questions on popular forecasting platforms[6]

Our report presents initial evidence that structured interviews of experts produce more informative questions about AI risk than the highest-engagement questions (as measured by unique users) on existing forecasting platforms.

Using predictions made by superforecasters (n=8), we compared the status quo questions to a subset of the AICT questions.[7] Most of the AICT questions (nine of 13) scored higher on VOI than all 10 status quo questions.[8]

VOI is based on each respondent's *expected update* in their belief about the ultimate question, not on how much a participant would update if an event happened. That is, it takes into account how likely the forecaster believes an event is to occur. If an event would result in a large update to a participant's forecast, but is deemed vanishingly unlikely to occur, it would have a small VOI. If an event would result in a large update, and is also considered likely to occur, it would have a high VOI.

Table E.1 compares the top five AICT questions to the top five status quo questions, as measured by superforecasters' ratings of a standardized metric of informativeness, which we call "Percentage of Maximum Value of Information" (POM VOI).[9] In this table and throughout the report, we refer to questions by their reference numbers. For a full list of the AICT questions and status quo questions selected from forecasting platforms by reference number, with operationalizations and additional information, see Appendix 1.

Question (See Appendix 1 for details)	Mean POM VOI
AI causes large-scale deaths, ineffectual response (CX50)	6.34%
Administrative disempowerment warning shot (CX30)	3.55%
Deep learning revenue (VL30)	1.68%

[6] By "informative," we mean that knowing the answer to one of these questions would make a larger difference, in expectation, to a participants' forecast of the ultimate question, in this case, "Will AI cause human extinction by 2100." For more on informativeness and the metric we use to assess it, see the section on Value of Information (VOI). Forecasting platforms are generally focused on making accurate predictions by aggregating many people's forecasts and usually allow participants to choose which questions to forecast. The questions that are popular on forecasting platforms are often questions that are important in themselves, more than as indicators of other events, and the platforms are not deliberately attempting to find high VOI questions.

[7] For more on the question filtering process, see Section 2.2.

[8] The four lowest-scoring AICT questions – EX50, HS50, NG30, and EX30 – ranked 12th, 13th, 14th, and 20th out of 23, respectively.

[9] At the time of data collection, we had not yet developed the POM VOI metric, so participants were not deliberately optimizing for it. Later, we found that POM VOI captured the idea of question informativeness better than VOI alone, which yields a number that is hard to interpret and contextualize. For a full list of questions analyzed, see Table 3.1.3. A comprehensive explanation of the POM VOI metric can be found in Appendix 4.

Power-seeking behavior warning shot (ZA50)	1.59%
Extinction-level pathogens feasible (CQ30)	1.37%
Superalignment success (STQ205 / STQ215)	0.28%
Kurzweil/Kapor Turing Test longbet (STQ9)	0.27%
Brain emulation (STQ196)	0.23%
Human-machine intelligence parity (STQ247)	0.14%
Compute restrictions (STQ236)	0.13%

Table E.1. *Ratings of how informative AICT questions are relative to status quo questions. The cells that contain status quo questions are highlighted in blue.*

Focusing on questions resolving in the near-term (by 2030), we found that questions generated with the conditional trees method were, on average, nine times more informative than popular questions from platforms ($p = .025$). While we did not find a statistically significant result for questions resolving in 2050-2070, in our sample AICT questions were still eleven times more informative on average. (More on VOI comparison)

Questions generated through the conditional trees method emphasized different topics than those on forecasting platforms

We also analyzed the extent to which questions taken from existing forecasting platforms effectively captured the topics raised in our expert interviews. We found that some topics (such as AI alignment-related questions and questions related to concrete AI harms) were of substantial interest to experts but had not received proportional attention on existing forecasting platforms, and that questions generated by the conditional trees method were meaningfully different from those taken from existing forecasting platforms.

The table below compares the topical distribution of the AICT questions to the status quo questions. (More on question uniqueness)

Category	AICT question set	Status quo question set
Social / Political / Economic	24% (29)	33% (131)
Alignment	20% (25)	12% (47)
AI harms	20% (25)	7% (27)
Acceleration	36% (44)	48% (191)

Table E.2 *Proportion of total questions that fell into each category; numbers in parentheses are total questions per category. While some questions fell into multiple categories (and thus proportions in each column should sum to more than 100%), proportions have been normalized for ease of comparison.*

We found weak evidence that superforecasters and experts value different types of questions

Given the small sample sizes involved, we are reluctant to make confident claims about the significance of the difference between the opinions of the superforecasters and the experts. However, we do see these results as providing prima facie evidence about which questions are the most informative for each group when making updates on the probability of AI-related extinction.

Our most notable finding when comparing the views of the superforecasters to those of the experts was that the superforecasters tended to value questions that focused on concrete harms caused by AI, rather than the experts' preference for questions regarding advanced AI capabilities or whether AI had been successfully aligned. (More on AI risk takeaways)

Figure E.1 shows examples of how experts updated on the ultimate question conditional on three of the highest-VOI indicator questions.

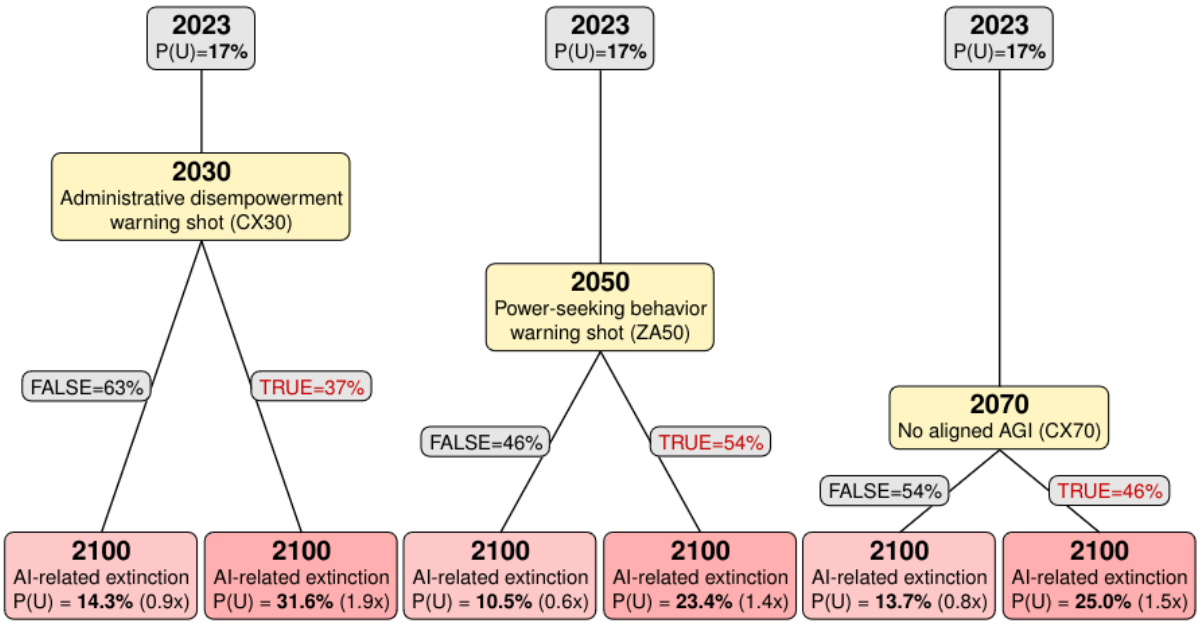


Figure E.1 *A diagram showing how experts update on three relatively high-VOI questions for different resolution years that scored particularly well on our VOI metric. Since experts answered different sets of questions, we derived P(U|C) and P(U|~C) (the probabilities on the bottom level) by multiplying the whole expert group's average P(U) of 17% by the average relative risk factor for each crux.*[10]

The table below provides more detail on each of the questions in the previous figure.

[10] Careful readers will note that the probabilities in this figure do not yield the mean POM VOI values we report (see Table E.1). Mean POM VOI tells us how valuable a crux is for a group, on average, by computing POM VOI at the individual level and then aggregating. The average relative updates, across individuals in the same group, sometimes tell a quite different story.

Question	Summary
Administrative disempowerment warning shot (CX30)	This question resolves positively if any of the following occur by 2030: AI systems in a country freeze assets affecting ≥1% of the population or ≥0.5% of GDP for over a year; prevent ≥5% of voting-age citizens from participating in a national election; restrict >3% of residents' movement for 3+ months; limit >10% of the population's internet access for 3+ months; or legally restrict >5% of citizens' access to certain housing or employment for 1+ year.
Power-seeking behavior warning shot (ZA50)	This question resolves positively if, by 2050, AI developers must attempt to disable or destroy an AI (costing >$5M) after it exhibits power-seeking behaviors such as acquiring significant resources, controlling financial accounts, manipulating humans, infiltrating key systems, seizing control of weaponry, self-propagation, or engaging in hacking.
No aligned AGI (CX70)	This question resolves positively if there is no AGI system which meets several “alignment” criteria by 2070. Those criteria include AGI-like capabilities (high question answering skill on hard questions in diverse subjects, high programming ability) and alignment (performing well on SOTA alignment tests, and expert surveys indicating that experts believe that AI is aligned).

Table E.3. *Example summaries of questions that experts found to be particularly informative.*

The conditional trees method still has disadvantages

While this case study suggests that the conditional trees method can generate informative forecasting questions, a primary limitation of the method as implemented is its high labor cost. The process involved conducting more than 20 interviews with subject matter experts, writing 75 forecasting questions, and eliciting conditional forecasts. In future work, we expect it would typically be more efficient to elicit fewer indicators within a conditional tree and to operationalize only 1-2 forecasting questions per interview before eliciting forecasts. The intensive process described in this case study would be most appropriate for particularly high-value topics with large pools of resources for research. Additionally, it may be possible to use LLMs or incentivized crowdsourcing for the question generation or filtering stages, making the process cheaper and less labor intensive. (More on limitations of our research)

Key takeaways

1. Preliminary evidence suggests that the conditional trees method of generating forecasting questions can result in questions that perform better on “value of information” metrics than popular questions on existing forecasting platforms.
2. The conditional trees method produced questions with a markedly different distribution of topic areas compared to those on existing forecasting platforms. Notably, the conditional trees approach led to a greater proportion of questions focused on AI alignment and potential AI harms, reflecting that certain expert priorities may be underrepresented in existing forecasting efforts.

3. In our limited sample, experts tended to find questions related to alignment and concrete harms caused by AI to be the most informative. Superforecasters also found questions relating to concrete AI harms to be informative, but were less likely than experts to find questions relating to alignment to be informative.
4. The conditional trees method as implemented in this case study is particularly labor intensive. We expect the most broadly useful versions of this process would take the underlying principles and 1) apply them to shorter interviews with smaller numbers of forecasting questions to operationalize, 2) leverage LLMs for elicitation and synthesis, and/or 3) utilize crowdsourcing at the question generation and filtering steps.

Key outputs

In addition to the above takeaways, we highlight key outputs from the report: the tangible resources developed during the course of the conditional trees process which we believe may be useful to others interested in replicating parts of the process.

1. We created a guide and replicable process for using conditional tree interviews to generate informative forecasting questions (see Appendix 6). This process can be implemented by organizations and individuals that need high-quality, informative questions.
2. We provide details of relevant metrics (e.g., '"Value of Information") that can be used to assess how informative each generated question is. See our public calculator for "value of information" and "value of discrimination" here.
3. In total, the conditional trees process generated 75 new questions relating to AI risk. The full operationalizations and resolution criteria of these questions are available in Appendix 1 of this report. We have posted several of the highest-VOI questions to two forecasting platforms and encourage interested readers to submit their own predictions. (See Appendix 7 for links)
4. We used our question metrics to create aggregated conditional trees that visually summarize the most important AI risk pathways according to small samples of experts and generalist forecasters. These aggregated trees can be found here.

Limitations of our research

Limitations of our research include:

1. The total number of participants in this study was small (n=8 forecasts on most questions, 24 interviewees to generate questions).
2. The forecasting tasks in this study were unusually difficult, involving low probability judgments, long time horizons, conditional forecasts, and "short-fuse forecasts" made very quickly.
3. Participants were all either experts who are highly concerned about existential risks from AI or superforecasters who are relatively skeptical, so we are not able to separate differences caused by risk assessment from differences caused by forecasting aptitude, professional training, or other factors.
   (More on limitations of our research)

Next steps

Further research related to this topic could include:

1. Studies on the same questions with larger numbers of forecasters, including by integrating the questions into existing forecasting platforms.
2. Replicating the conditional trees process in domains other than AI risk.
3. Following up as questions begin to resolve in 2030 to assess whether forecasters update their views in accordance with their expectations.
   (More on next steps)

Glossary

AI Conditional Trees (AICT) question set
The set of questions generated by the AI conditional trees process described in this report.

Conditional tree
A simplified Bayesian network, in which each node is an event that may or may not occur, and each connection between nodes has the factor by which the next node is more or less likely if that one happens. In this report, the conditional trees ultimately ask how likely it is that AI causes human extinction by 2100, and each node is an event that affects the likelihood of that ultimate outcome.

Operationalization
The process of making a question about a future event into a resolvable forecasting question. For example, if a prompt said “there is major progress in interpretability by 2030” the operationalized question would contain a specific way to resolve that question so that there can be no future dispute about whether the progress counts as “major.”

Percent of Max (POM)
When we present VOI for a question, we also present the percentage of the maximum VOI (POM VOI) it captured in order to contextualize the magnitude of the results. The POM VOI of a question can be interpreted as the fraction of the uncertainty about the ultimate question U the question resolves, in expectation.

Question prompts
General topics of questions that we then operationalized into forecasting questions. For example, “major progress in interpretability by 2030” could be a question prompt, although it is not a clearly resolvable forecasting question.

Short-fuse forecasts
Very quickly estimated forecasts, in which each participant spent no more than one minute per question and gave a snap judgment.

Status quo questions
Questions on AI that we selected from existing forecasting platforms on the basis of their popularity (largest number of unique users) and other criteria. See 2.3 Selection of status quo questions.

Ultimate question / Ultimate outcome (U)
The “ultimate question” that all of the intermediate questions help predict. In this study: “Will AI cause human extinction by 2100?”

Value of information (VOI)
VOI is a measure of how much knowing the answer to a question would change an individual's belief, in expectation. This is useful for understanding why individuals believe what they believe and what would change their minds.

1. Introduction

For policymakers to use forecasting in their work, they need accurate forecasts, but—perhaps equally important—the forecasts need to be about decision-relevant questions. Knowing which questions will be the most valuable to forecast on can be difficult. How can policymakers identify the short-term events that are most relevant to important long-term outcomes?

Here we present a tool, the conditional tree method (figure 1.1.1), which can distill complex issues into a few key uncertainties. We apply it to a topic of increasing public concern: "Will advanced artificial intelligence pose an existential threat to humanity in the 21st century?" Using a specialized interview process, we learn what subject matter experts believe are the best warning signs for this risk in the coming decades. Then we use metrics based on conditional forecasting to quantitatively measure the relevance of these warning signs. This allows us to winnow down to a few highly relevant indicators of increased risk to humanity from AI.

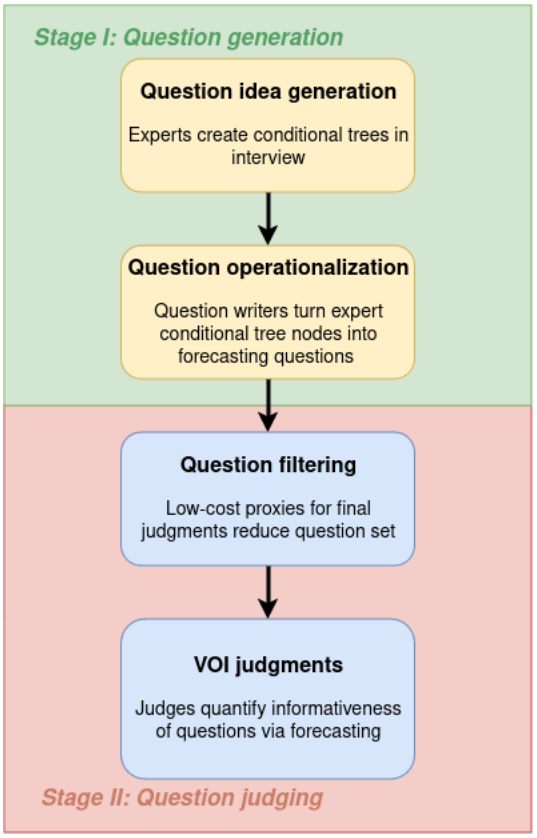


Figure 1.1.1: *The conditional trees process*

The conditional trees approach[11] represents a new set of priorities in the field of forecasting. Most previous forecasting research focused almost exclusively on identifying accurate forecasters and improving forecasting accuracy. But comparatively little work was invested in choosing forecasting targets. In order to mature into a practically applicable body of knowledge, the field must look beyond optimizing forecasts and toward optimizing the questions we ask.

1.1 A method for generating and judging high-value questions

Some forecasting tournaments and platforms have already begun to utilize domain experts to generate questions with real-world relevance. However, many of these efforts are

[11] Several related methods, such as Delphi and Bayesian Network elicitation, may be useful to forecasting research in similar ways. See Bernice B. Brown, "Delphi Process: A Methodology Used for the Elicitation of Opinions of Experts," Rand Corporation report (September 1968) and Judea Pearl, *Probabilistic Reasoning in Intelligent Systems*, (New York, Morgan-Kaufman: 1998).

relatively *ad hoc*, producing inconsistent results and plausibly missing many high value forecasting targets.

For example, for the Existential Risk Persuasion Tournament (XPT),[12] the question preparation phase enlisted domain experts to comment on the prospective question set in a relatively unstructured way. While this undoubtedly improved the question set, it did not identify the most informative questions within the set.

To leverage the expertise of domain experts more fully, we propose a more in-depth, systematic approach: expert elicitation structured around conditional trees.

Why conditional trees?

Conditional trees represent beliefs through a tree-like structure, using nodes to represent events that influence the probability of an ultimate outcome. In the tree in Figure 1.1.2, for example, if you know someone is vaccinated, they are half as likely to be infected than if you were unsure whether they were vaccinated. Then, if you know they have been exposed, they are 3.5x as likely to be infected.[13]

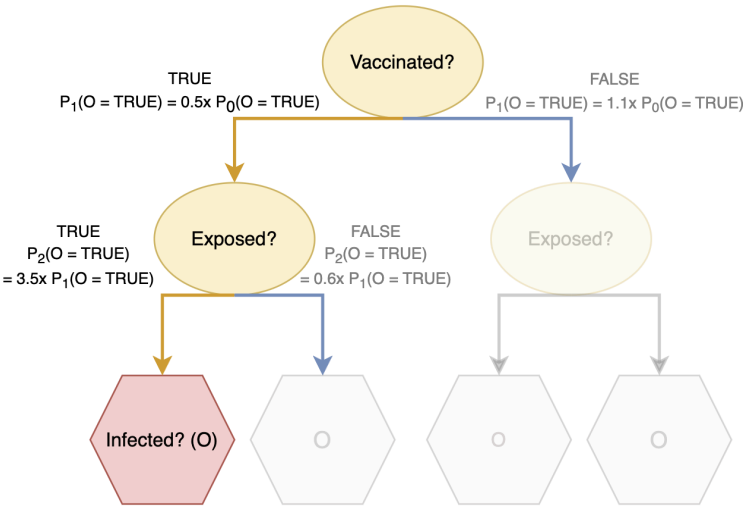


Figure 1.1.2: *Example conditional tree diagram*

In this study, the ultimate outcome was the probability of extinction due to AI by 2100, and the nodes are events that make that outcome more or less likely. The tree structure makes the conditional probabilities beneath a forecast explicit and visible, and may help forecasters narrow in on specific, important factors.

Participants initially provided an estimate of the probability of AI-related extinction by 2100 (the "ultimate question"), represented by O in Figure 1.1.2. Interviews then focused on identifying key indicators on the pathway to AI-related extinction. Participants selected two to five indicators for deeper analysis to understand how they might alter the risk of AI-related extinction. These factors then became the antecedents in the tree: for each of the indicators selected to be included in the tree, participants gave forecasts for how much their forecast of the ultimate outcome would change if that event happened.

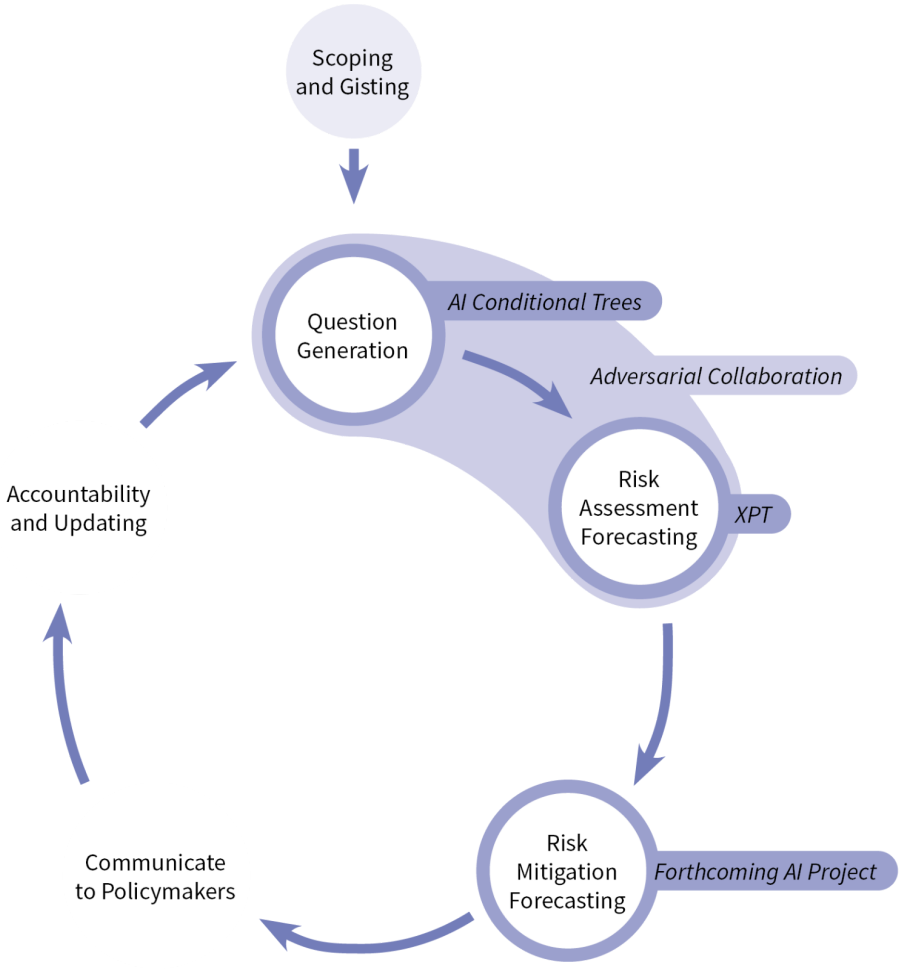
The ultimate outcome (for our purposes, the probability of extinction due to AI by 2100) is an important parameter: the rest of the network's relevance cascades from the outcome. But provided we're able to identify an outcome with strong bearing on present policy decisions, we can ask experts to decompose the intervening time into possible events which would reflect a greater or lesser likelihood of reaching that outcome. Thus, these intervening events

[12] Karger et al., "Forecasting Existential Risks: Evidence from a Long-Run Forecasting Tournament," 2023. https://forecastingresearch.org/xpt (a) (XPT report).
[13] These numbers are intended to be illustrative and are not based on actual vaccine data.

must themselves possess policy-relevance, in proportion to the strength of their relationship with the outcome, and the likelihood of observing them.

Conditional trees are a type of Bayesian network (BN).[14] BNs explicitly represent probabilistic relationships between outcomes and their antecedents.[15] This structure encourages experts to generate maximally relevant antecedents, and also provides us with a framework for measuring question relevance. But unlike some other forms of BNs, conditional trees are a relatively easy tool to learn. In our study, interviewees were able to grasp the necessary basics in around 10 minutes. This means that conditional trees may be more practical for interviews with subject-matter experts, who may not be experts in statistics or other domains that more often use BNs.

How does the conditional trees method fit into the forecasting research process?



[14] Judea Pearl, "From Bayesian Networks to Causal Networks," in *Mathematical Models for Handling Partial Knowledge in Artificial Intelligence*, ed. Giulianella Coletti et al., 160. Boston, MA: Springer, 1995. https://doi.org/10.1007/978-1-4899-1424-8_9 (a)

[15] This relationship can be causal, but it does not need to be; in this project we did not constrain conditional trees to only causal relationships, nor did we probe expert models for causality in the interviews.

Figure 1.1.3: *Life cycle of an impactful forecasting project.*

The AI Conditional Trees project is an in-depth investigation into how to generate informative forecasting questions. Question generation is the first step in the life cycle of an impactful forecasting project, illustrated in Figure 1.1.3.

Many earlier forecasting research projects have focused on identifying the most accurate forecasters and on improved methods for aggregating their forecasts. But to be useful to decision makers, forecasting research must move beyond those questions and incorporate forecasting into a process that includes question generation, considering actions based on forecasts, communicating with policymakers, and generating new questions.

Before the cycle starts, we begin with "scoping and gisting," in which we consider the questions we want to answer, the scope of the possible project, and the general arguments ("gists") on each side. We then begin the cycle by generating questions, through processes like the AI conditional trees method, aiming to find the forecasting questions that would be most informative to decision makers. Next, we elicit forecasts on those questions, to assess risk and understand which potentially dangerous events are most likely and in what circumstances. We then elicit "risk mitigation forecasts," asking experts and skilled forecasters to predict which policies would most decrease risk and what the costs might be for implementing them.

Once we have completed these stages, we communicate that information to policymakers, and ask them whether it is useful and what would make it more relevant to their work. Their feedback gives us more information we can use for the next stage of question generation, and we begin the cycle again.

The cycle as depicted is somewhat stylized, and many forecasting projects will not include all of these stages. But thinking of AI conditional trees in the context of the "forecasting life cycle" helps us contextualize this work and think about how to incorporate it into our future research.

Measuring question value

In order to form the feedback loop necessary for a dramatic improvement in the decision-relevance of forecasting questions, we need a means of quantitatively measuring the value of a forecasting question.

Policymakers' actions are often guided by a few important questions in their domain, like "What will be the effects of climate change over the next century?" or "Will our economy remain competitive in the world in the long-term?" Such questions are difficult to resolve because they refer to the distant future, and they may also be relatively complex or difficult to specify clearly. But often one can find nearer-term antecedent questions which are easier to resolve, and which would reduce some uncertainty about the "ultimate" question. For example, in a study forecasting the effects of climate change, with the ultimate question, "Will more than 2 billion people die or be displaced due to climate change by 2100?," the question "What will the average global temperature be in 2040?" might be a good antecedent question. It would not give a forecaster the full answer to the main question, but

knowing what the global surface temperature will be in 2040 would be at least somewhat helpful for forecasting the effects of climate change by 2100.

Thus, one way of conceptualizing the value of a forecasting question is to ask, “How would the answer to this question affect our expectation about an ‘ultimate’ question we care about?” There are several distinct ways of expressing this mathematically, which we collectively refer to as “Value of Information (VOI).”

Conceptually, VOI measures how important a potential crux question (“C”) is to a participant’s forecast of the ultimate question we care about (“U”, in this case: AI extinction risk by 2100), in expectation. That is, how much would a participant update on AI extinction risk by 2100 based on whether a crux happens, weighted by how likely that crux is to happen. A high VOI question for a given participant will therefore be one that a) that participant thinks has a meaningful chance of happening and b) meaningfully affects that participant’s forecast on the ultimate question.

VOI is a useful metric for understanding why individuals believe what they believe and what would change their minds. A technical explanation of VOI can be found in Appendix 4. To build intuition for using the VOI metric, we provide this calculator (a) in which users can input their own values. We also provide a more comprehensive R software package (a) for calculating it.

2. Methods

2.1 Question generation

Sampling interviewees

Our sample included 24 interviewees in total: 21 “expert” interviewees, and 3 “superforecaster” interviewees. We aimed to include in our sample representatives of four quadrants of a strategically important belief space (see Figure 2.1.1):
1) short timeline for AI progress, high estimated risk from AI;
2) short timeline for AI progress, low estimated risk from AI;
3) long timeline for AI progress, low estimated risk from AI; and
4) long timeline for AI progress, high estimated risk from AI.[16]

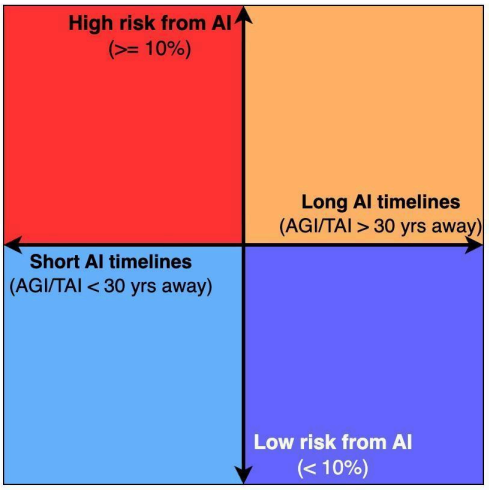


Figure 2.1.1: *Target groups for sampling*

We gathered our expert sample via snowball sampling, seeded from recommendations from our funders and our networks. We do not expect our interview sample was particularly representative of any given group, such as AI experts. The goal of this project was to develop the trees process and assess whether it led to higher value questions, which did not require a representative expert sample. Our superforecaster sample was taken from the set of superforecaster participants in the Existential Risk Persuasion Tournament (XPT)[17] who had shown particularly high engagement. Candidate interviewees were approached for interview with a monetary incentive for producing the “highest value” questions in our interview-derived question set.

[16] We defined “high risk” as forecasting >10% chance of extinction due to AI by 2100, and low risk as <10%. We defined “long AI timelines” as forecasting >30 years until transformative AI or artificial general intelligence and “short AI timelines” as less than 30 years.
[17] Karger et al., XPT Report.

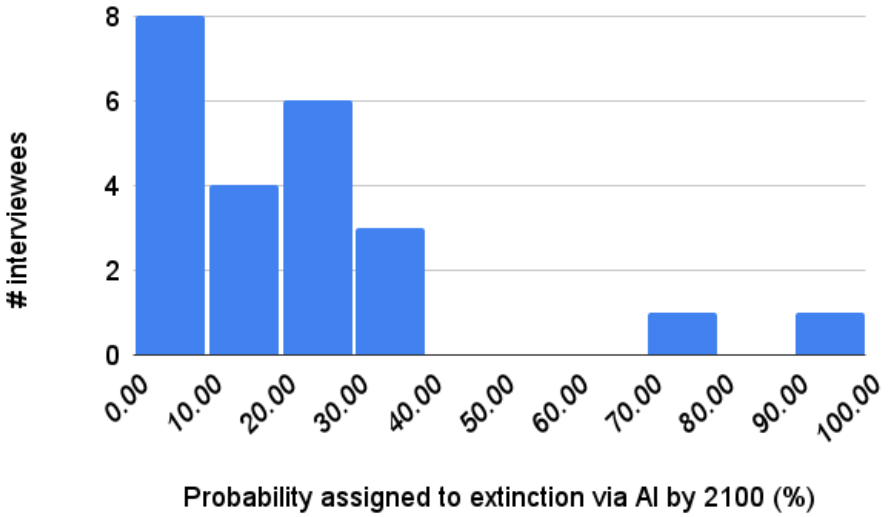


Figure 2.1.2: *Histogram of interviewees' original forecasts of probability of extinction via AI by 2100.*[18]

The majority of our expert sample had academic or professional experience pertaining directly to AI risk, such as experience in technical AI safety or AI governance (13/21 expert interviewees). Others were included for having publicly expressed views on AI risk indicating a high level of engagement with the topic and having expertise in a complementary field, such as machine learning (7/21 expert interviewees). Finally, a small number of our expert sample had expertise in a complementary field, but had not expressed detailed views on AI risk in public (2/21 expert interviewees). Most of our expert sample held senior positions within their fields, as professors, directors of organizations, leaders of research teams, or similar (13/21 expert interviewees).

Our expert sample skewed toward the top left quadrant in figure 2.1.1, "high risk/short timelines." Of 21 expert participants, 13 estimated the risk of extinction from AI by 2100 to be >10%. Only one of our expert sample estimated the risk to be <1% by 2100, whereas the median expert in the XPT predicted 3%. Although we did not solicit AI progress timelines directly from interviewees, interview content generally suggested a positive relationship between beliefs in increased risk and shorter timelines in our sample.

Because of this skew in our expert sample, we chose to ensure some representation of the bottom two quadrants in figure 2.1.1 (low risk from AI) by selecting three superforecaster interviewees who forecast <10% probability of extinction from AI by 2100.

[18] One interviewee is not represented in this graph because in the interview, they responded ">0.1%, <50%" rather than give a point estimate.

Interview process

Interviews were 1-on-1, ran for roughly 60 minutes and followed a semi-structured format. By default, interviews aimed to trace one plausible path of increasingly strong signals of heightened AI risk at three successive timepoints before 2100.[19] Interviewers[20] were allowed some latitude for individual approaches, but generally followed this basic structure:[21]

1. Introduction, task instructions
2. Elicitation of P(AI-related extinction by 2100)
3. Node generation
4. Wrap-up questions

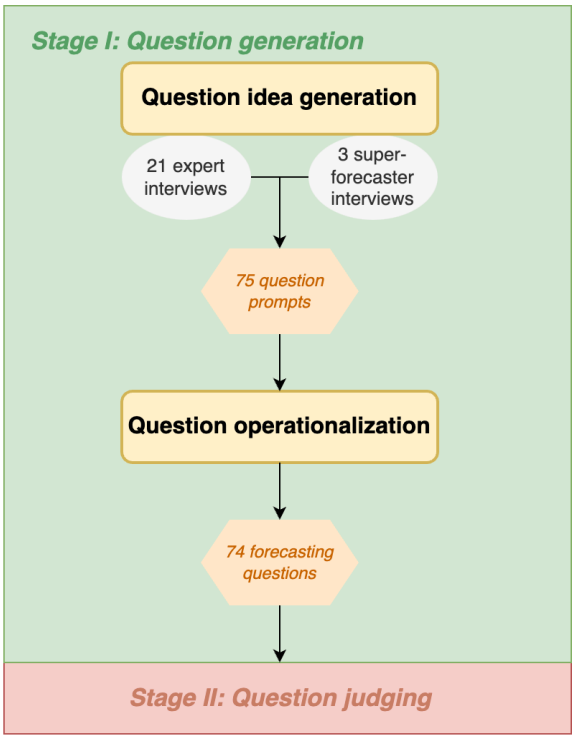


Figure 2.1.3: *The conditional tree workflow (I)*

Interviewees were first given a very brief summary of the aims of the project, a short explanation of conditional trees, and a statement of the goals of the interview. Interviewees were also told that they would be awarded $1,000 if a forecasting question derived from their interview was one of the "highest value" forecasting questions generated by the project.[22] This introductory section of the interview typically took 10 minutes or less.

Next, interviewees were asked to give their best guess probability for the project's "ultimate question," namely "AI-related extinction by 2100," which was operationalized as in the 2022 XPT.[23] Following the probability elicitation, we sometimes asked participants warm-up questions, for instance asking them to name possible "driving forces" influencing their views.

Interviewees would then begin the node generating phase of the interview, which comprised the majority of interview time. Although we began the project with a set of three predefined

---

[19] Occasionally interviews instead focused on signals of decreased AI risk, at the discretion of the interviewer (e.g., if the interviewee's starting risk estimate was extremely high).
[20] Interviewers were: Tegan McCaslin (11/24 interviews), Josh Rosenberg (10/24 interviews), and Ezra Karger (3/24 interviews)
[21] For a full description of the interview process, see Appendix 6.
[22] This incentive was not explained in further detail given time constraints of the interview.
[23] In the XPT, "Extinction" was defined as "reduction of the global population to less than 5000," and extinction was considered "due to AI" if AI was the direct or proximate cause of the deaths. This definition encompasses events that would not have occurred or would have counterfactually been extremely unlikely to occur "but for" the substantial involvement of AI within one year prior to the event. For more details, see Karger et al., XPT Report, 134.

For some interviewees, this was a question for which they had already devoted substantial time (in the XPT or other contexts) forming a quantitative forecast, and thus such participants were able to offer a relatively quick probability judgment. Most participants had previously spent substantial time thinking about the possibility of AI-related extinction, but not as much time forming a precise quantitative estimate for the date in question, and many expressed hesitancy about their answer in the interview.

years to ask participants about (2030, 2050 and 2070),[24] it soon became clear that this was not the best choice of years for participants with short AI progress timelines. Therefore, we began in the node generating phase to ask participants to propose a suitable set of years for their own trees (see Appendix 3 for the distribution of years chosen).

For each node, we took interviewees through a process of brainstorming, selection, and fleshing out. We would then elicit a probability of AI extinction by 2100 conditional on the node. We will refer to these pre-operationalization nodes as **question prompts.**

Interviewers took detailed notes, and most interviews were recorded (with participants' permission). Further details on interview technique can be found in Appendix 6.

Operationalizing question prompts as forecasting questions

Question prompts were generally not fully resolvable forecasting questions, though some were operationalized in more detail than others. We considered it an inefficient use of interview time to focus on constructing forecasting questions with detailed resolution criteria, and also not the comparative advantage of expert interviewees generally. Instead, an internal question writing team[25] turned question prompts into fully operationalized forecasting questions, with the help of notes from the interview and feedback from the interviewer.

The primary goals of question writing in this project were:

1. To capture as much of a question prompt's original intent as possible, while still making questions highly resolvable.
2. To optimize the value of information from the question by adjusting thresholds or removing elements which made the probability of a positive or negative resolution too extreme.

We developed a template for the question writing process, which encouraged question writers to first consider multiple distinct ways the interview node could be operationalized. They then analyzed these options with respect to several important criteria:

- How much the question captured the most relevant aspects of the original interview node;
- How efficiently the question captured relevant aspects of the original interview node;
- Salient hypothetical cases of false positive resolution and false negative resolution;
- How clear cut or practically feasible resolution of the question would be;
- Amount of cognitive load for forecasters.

The question writer and reviewer would then jointly decide which formulations to include in the final question on the basis of these criteria. Finally, a more detailed set of resolution

[24] These resolution years were chosen to match XPT questions.
[25] Question writers were Tegan McCaslin, Taylor Smith, Josh Rosenberg, Rose Hadshar, Adam Kuzee, Ezra Karger, Arunim Agrawal, and Bridget Williams. One primary question writer was assigned to each question prompt, and would draft several different versions of the question, using the interview notes as an aid to understanding the interviewee's underlying models. These drafts would receive feedback from the rest of the question writing team, and in particular from the relevant interviewer. This interviewer had final say over revisions and finalizing the question.

conditions would be written and incorporated into a “conditional tree summary document”, which could then be sent to the interviewee for feedback.

2.2 Judging questions and constructing aggregate trees

The question generation phase yielded 75 questions, some of which were very similar to one another, so our next task was to filter them and select the most useful questions to construct conditional trees. We began by eliciting “short-fuse” forecasts on each question, in which forecasters spent about one minute per question giving quick judgements that allowed us to estimate a rough VOI for each question. For the thirteen questions that passed this initial screen, we conducted a longer survey, asking participants to spend more time forecasting how likely each question is to resolve positively and how much difference it would make to their ultimate forecast of the likelihood of extinction due to AI by 2100.[26]

Because participants in this study were all either (i) superforecasters who forecasted less than 1% likelihood of extinction due to AI by 2100 or (ii) people with professional AI risk-related experience who forecasted more than 1% likelihood of extinction due to AI by 2100 (with one exception, they forecasted at least 5%), we targeted these two socio-ideological camps separately in our question rating. We denote these groups, respectively, as “skeptical superforecasters” and “concerned experts.”

First pass filtering of the question set

Our full set of operationalized nodes included 75 questions, many of which were relatively overlapping. It would have been inefficient and excessively cognitively taxing to participants if we had attempted to elicit full 20-minute VOI judgments on each of the 75 questions. Therefore, we performed a first-pass filter on the question set using “short-fuse” forecasts.

We elicited VOI judgments in a “short-fuse” format from 8 skeptical superforecasters. This required very quick judgments, approximately 1 minute per question.[27] Separately, we also collected question data from a set of 5 “concerned expert proxies,”[28] asking them to rank order the question set and provide VOI judgments for a subset.[29] However, this method may have been substantially flawed, as actual experts did not ultimately think the questions selected by the proxies were more informative than other questions.

---

[26] The initial screen was not simply a VOI threshold. To get a diverse question set, we wanted to include at least one question from each of the following categories: 1) high VOI for superforecasters, 2) high VOI for experts, 3) high VOD between experts and superforecasters, 4) jointly high VOI between superforecasters and experts, 5) randomly chosen representative of the bottom half of the AICT question set, and 6) top comparable question from outside the AICT set. Choosing cutoffs separately for each of these categories resulted in thirteen questions.

[27] Participants gave estimates for the probability of the question resolving positively (P(c)), and the probability of AI extinction *conditional* on the question resolving positively (P(U|c)). We then used these figures to calculate each respondent’s VOI for each question.

[28] The “concerned expert proxies” were teammates or collaborators who had had extensive contact with concerned experts, who we expected to be able to model this group’s views well.

[29] Instead of giving probability judgments on all 75 questions, the concerned expert proxies chose and rank-ordered their top 10 questions from each of: the set of first-tier nodes (usually 2030); the set of second-tier nodes (2035-2050); and the set of third-tier nodes (2040-2070). They then provided short-fuse VOI judgments for only the questions they had ranked in their top 10 for each position.

For superforecaster data, we ranked questions according to median VOI in the filtering round.[30] The filtered question set included thirteen questions including seven questions for the first tier (dates up to 2030) and six questions for the second tier (2031-2070).[31]

Main question-rating survey

After the initial filtering, we further refined our question set using surveys, in which *skeptical superforecasters* and *concerned experts* were asked for more detailed forecasts on the filtered question set. We offered a fixed sum as an incentive for survey completion. Superforecasters answered a longer survey containing all thirteen questions. Because of experts' time constraints, each expert answered a shorter survey containing a random subset of the questions.

The main survey superforecaster sample (n=8) was the same as the filtering survey

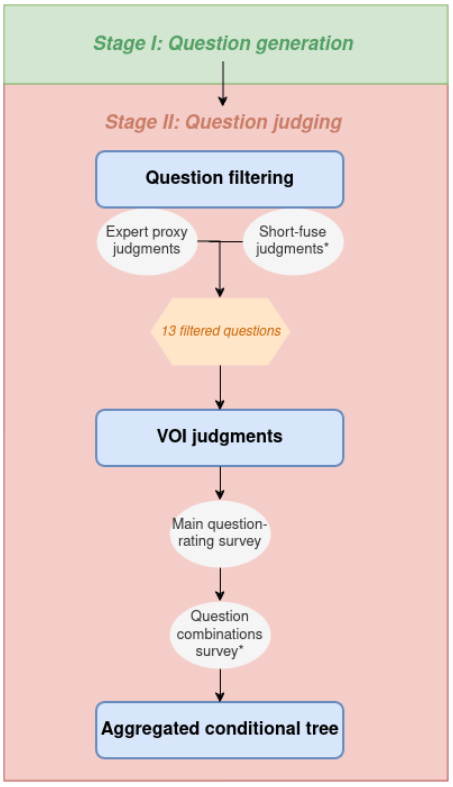


Figure 2.2.1: *The conditional tree workflow (II)*

**Denotes stages which only superforecasters participated in.*

[30] For the concerned expert-proxy data, we ranked questions via ranked choice voting. We also employed the value of discrimination (VOD) metric, which measures the change in disagreement between two forecasters a question is expected to make (see Appendix 4). VOD was determined by the median of pairwise VOD across both skeptical superforecasters and concerned expert-proxies. We excluded questions which closely resembled other questions ranked higher, those which the question writing team did not operationalize, and those with the lowest individual-level VOI ranking.

[31] The filtered question set included the following questions. See Table 3.1.3 for concise question summaries.

Node 1 (dates up to 2030):

- CQ30, VL30, NG30, respectively: The VOI top-ranked node for skeptical superforecasters; the VOI top-ranked node for concerned expert proxies (also ranked 2nd for VOD); the VOI 2nd-ranked node for concerned expert proxies;
- CX30: The top-ranked node for VOD;
- ZD30: Included for having relatively good agreement on high VOI between groups;
- EX30: Randomly chosen from the set of nodes ranked in the bottom half by both groups, as a check on the validity of the filtering process;
- STQ9: A question from outside our question set, the most-upvoted AI question on Metaculus resolving around 2030.

Node 2 (2031 - 2070):

- ZA50, EX50, VL70, respectively: The VOI top-ranked node for skeptical superforecasters; the VOI top-ranked node for concerned expert proxies; the VOI 2nd-ranked node for concerned expert proxies;
- CX70: The top-ranked VOD node;
- HS50: Randomly chosen from the set of nodes ranked in the bottom half by both groups;
- STQ247: The most-upvoted AI question on Metaculus resolving post-2030.

sample. At this point, the sample had also participated in a lengthy adversarial collaboration with a camp of AI-risk concerned experts.[32] Thus they had spent significant time developing their own beliefs on the topic and engaging with opposing beliefs.

The expert sample (n=11) was drawn from the candidate participant list from the AI adversarial collaboration.[33]

Superforecaster survey

In the superforecaster survey, we presented all 13 questions of the filtered question set in Qualtrics, shown in two parts, first 2030 questions and then 2050-2070 questions. Within each part we randomized question order. Participants were instructed to spend approximately 20 minutes per question, to give their own beliefs, and separately to estimate the beliefs of the concerned expert group.

We first asked for (1) each participant's own forecast of the probability of AI-related extinction by 2100 and (2) each participants' forecast of what experts would forecast about the probability of AI-related extinction by 2100.[34]

We then asked participants for forecasts on each of the 13 questions from the filtered question set. Each forecasting question contained moderately detailed resolution criteria, as well as links to reference information where possible. In the survey, answers were checked for logical coherence, and respondents were prompted to revise if necessary.[35] At the end of each part, we gave participants the opportunity to review all questions and answers from that section and revise if they wished.[36]

A supplementary survey using the same protocol as above with questions drawn from the "status quo" question set (questions from forecasting platforms (see Appendix 3.2) was administered at a later date. This survey also included two further questions from the AI conditional tree set which had initially been eliminated in the filtering stage.[37]

---

[32] The eight superforecasters in this sample took part in FRI's Adversarial Collaboration project that brought together generalist forecasters and domain experts with divergent views on AI's long-term risks to humanity. See Forecasting Research Institute, *Roots of Disagreement on AI Risk: Exploring the Potential and Pitfalls of Adversarial Collaboration* (2024) (a).

[33] For more detail on the selection pool, see Roots of Disagreement on AI Risk. The "AI-concerned" expert for this project consisted of domain experts referred to us by our funder and the broader effective altruism community.

[34] Respondents were not able to revise this forecast later in the survey.

[35] Once forecasters submitted their answers on a question, the survey checked for coherence, and then prompted the respondent to revise their answers if the coherence condition was not met. Coherence requires that $P(U) > P(U|c)P(c)$, where $P(U)$ is the forecaster's probability of the ultimate question U resolving positively, $P(U|c)$ is the probability of U resolving positively if the crux c resolves positively, and $P(c)$ is the probability of the crux resolving positively. This coherence prompt was not repeated on a question if the respondent failed to give coherent revised answers on that question. For any answers which remained incoherent after the respondent finished the survey, we followed up and requested revision.

[36] Due to coding errors in an early version of the survey, not all participants were given an opportunity to review their answers in the survey. We instead asked such participants to manually review their answers afterward.

[37] The two questions added to the supplementary survey were HB30 and CX50. See Appendix 1 for full question descriptions.

Expert survey

Experts were given the choice of a long or short version of the survey, including 6 and 3 questions, respectively. Each respondent saw a random subset of the 13 filtered questions. Experts were asked only to provide their own beliefs, without forecasting superforecasters' beliefs. Apart from these changes, the survey was identical to the superforecaster survey.

Question combinations survey

Because individual question ratings are not sufficient to build a full conditional tree with multiple intermediate nodes, we followed up the main question-rating survey with a survey eliciting judgments for every combination of four top-scoring questions from the main question-rating survey. As this is a relatively sophisticated and labor-intensive task, we administered it only to our skeptical superforecaster sample.

This elicitation was conducted in a Google Sheets form, and included top-scoring questions (either by POM VOI or z-score[38]) as previously rated by this sample: CX30, CQ30, CX50, and ZA50. VOI judgments were elicited for each of the sixteen combinations of "yes" and "no" resolutions for each of the four questions (i.e., all resolve positively; CX30 resolves positively and the rest negatively; CQ30 resolves positively and the rest negatively; …; all resolve negatively).

See Appendix 5 for further survey details. The image below presents the elicitation format.

	A	B	C	D	E	F	G	H	I
1		Administrative disempowerment warning shot (2030)	Feasibility of extinction-level pathogens (2030)	AI-related deaths (2050)	Power-seeking behavior (2050)	P(scenario)	P(U|scenario)	*Comments (optional)*	
2	Starting **P(U):**	**P(event):**	**P(event):**	**P(event):**	**P(event):**				**This is your implied P(U):**
3	**XX%**	**XX%**	**XX%**	**XX%**	**XX%**				**0.0000%**
4		**P(U|event):**	**P(U|event):**	**P(U|event):**	**P(U|event):**				
5		**XX%**	**XX%**	**XX%**	**XX%**				
6									
7		**Scenario interactions:**							
8		1	1	1	1				
9		1	1	0	1				
10		1	1	1	0				
11		1	1	0	0				
12									
13		1	0	1	1				
14		1	0	0	1				
15		1	0	1	0				
16		1	0	0	0				
17									
18		0	1	1	1				
19		0	1	1	0				
20		0	1	0	1				
21		0	1	0	0				
22									
23		0	0	1	1				
24		0	0	0	1				
25		0	0	1	0				
26		0	0	0	0				
27						0.00%	**<-- This should sum to 100%**		

[38] For details on the selection criteria, see Section 3.2. A z-score indicates how many standard deviations an observation is from the mean and in which direction. David S. Moore, George P. McCabe, and Bruce A. Craig, *Introduction to the Practice of Statistics*, 6th ed. (New York: W. H. Freeman and Company, 2009), 61.

Figure 2.2.2: *Elicitation format for combinations (or "scenarios") survey. Superforecasters were asked to provide forecasts for each of the scenarios in the yellow cells.*

2.3 Selection of status quo questions

For comparison, we selected a set of pre-existing AI forecasting questions from popular forecasting platforms. Questions were restricted to those with dichotomous resolution which did not directly ask about AI causing human extinction. We selected questions with the largest number of unique users engaging with them, rather than by forecast or trading volume, which is more vulnerable to individual differences in updating frequency. We also restricted the number of questions written by known public figures (e.g., Scott Alexander, Eliezer Yudkowsky), as their outsized performance relative to other questions seemed primarily due to their personal following. For a later analysis regarding the distribution of question topics (see section 4.2 Distribution of question topics), we tagged these questions as "acceleration," "alignment," or "social/political/economic" using our judgment of their subject matter.

From Manifold Markets we selected three unique questions:

- STQ47 (2030 set) - Largest total number of traders (1023), tagged "acceleration"
- STQ149 (2030 set) - Largest number of traders for a non-public figure question (355), tagged "acceleration"
- STQ19 (2030 set) - Largest number of traders for a non-public figure question, tagged "social / political / economic"

From Metaculus we selected four unique questions:

- STQ196 (2050-2070 set) - Largest number of forecasters after those included in the main survey (424), tagged "acceleration"
- STQ152 (2030 set) - Next largest number of forecasters (325), tagged "acceleration"
- STQ232 (2050-2070 set) - Next largest number of forecasters for 2050-2070 set (263), tagged "acceleration"
- STQ236 (2050-2070 set): Large number of forecasters for a 2050-2070 question, tagged "social / political / economic"

We selected two questions found on both platforms:

- STQ9 (2030 set): Large number of forecasters/traders, tagged "acceleration"
- STQ215 / STQ205 (2030 set): Large number of forecasters/traders, tagged "alignment"

3. Value of information (VOI) results

In this section we present the results of a quantitative analysis of question quality for our expert-derived "AI Conditional Tree (**AICT**)" question set. We rate these questions using metrics which factor in conditional and unconditional forecasts from surveyed populations, and reflect the strength of the relationship between the question and a possible future outcome (here, "AI-related extinction by 2100.") We selected two groups for the survey—subject matter experts and superforecasters—analyzed separately.

These VOI results are presented in the spirit of a demonstration of methods, and we would caution readers not to place undue weight on the question ratings. Given the very limited number of survey participants,[39] the views captured here are unlikely to be representative of those of subject matter experts or skilled forecasters more generally. Furthermore, at the time of this report, eliciting conditional probabilities is a relatively new practice, and with many wrinkles still to be ironed out. Nevertheless, there are a few observations worth highlighting.

Among questions resolving in 2030, both groups rated "Administrative disempowerment warning shot" (CX30) as leading to relatively large updates on the probability of AI-caused extinction by 2100 in expectation: it ranked first with both groups for our main metric, POM VOI, and was relatively robust in sensitivity analysis. The question refers to a scenario in which *"AI systems are preventing a substantial number of people from accessing financial assets, voting, moving freely, accessing the internet, or accessing certain types of housing or employment."*

As with many of the questions in our sample which performed well, CX30 benefited from one strongly positive opinion. Views about question value, even within the groups, were highly heterogeneous, and for all questions there was at least one respondent who took little or no information from it.

In the rest of this section, we:

- Provide a summary of the methods, metrics, and terminology used in this analysis and explain how to read a conditional tree (More)
- Summarize the question informativeness ratings for superforecasters and subject matter experts (More)
- Present aggregated trees that show the most informative questions at each timepoint for both superforecasters and subject matter experts (More)
- Provide details on the value of information ratings for all forecasting questions we surveyed superforecasters and subject matter experts about (More)

Summary of VOI methods, metrics and terminology

We surveyed two groups: a) forecasters with a strong track record of short-term accuracy, who also estimated a relatively low chance of AI-related extinction by 2100 (**"skeptical**

[39] 8 superforecasters (7-8 respondents per question) and 11 domain experts (4-6 respondents per question).

superforecasters") (*n* = 8 total, 7-8 respondents per question); and b) subject matter experts in fields related to AI risk, who also estimated a relatively high chance of AI-related extinction by 2100 (**"concerned experts"**) (*n* = 11 total, 4-6 respondents per question).

Due to the high cost of obtaining forecasts on all 75 questions, we evaluate only a subset of questions (13 in total). These were selected for their performance in a preliminary filtering round, though our data suggests that this filtering round was a weak predictor of main question-rating survey results, especially for our expert sample.[40] We also include in our survey the most popular (as of July 2023) AI questions from Metaculus, one each for 2030 and for the time period 2050-2070.

For each forecasting question, we asked respondents for their probability that it would resolve TRUE, and for their probability that AI extinction by 2100 would resolve TRUE, conditioned on the forecasting question resolving TRUE. We use Kullback-Leibler VOI (**KL VOI**, or simply **VOI** from this point forward) as our VOI measure.[41]

We focus on the *percentage of the theoretical maximum VOI* (**POM VOI**, or simply **POM**) that a question achieves as our main result.[42] In some places we also report the z-score of a question's POM VOI value for a given respondent (**POM-z VOI**, or simply **POM-z**). This value is useful if you believe individual respondents may have a bias toward giving higher or lower answers in general, or toward reporting an overall wider range of VOI values. It is particularly useful in the case of the expert results, as each expert answered only a random subset of all survey questions, and thus the influence of individual response biases on the resulting rank order of questions is potentially problematic. We suggest interpreting POM-z as a robustness check on the main POM results.

We aggregate POM and POM-z over respondents using the arithmetic mean. This sometimes has the effect that a single extreme response dominates the aggregate; however we believe this is appropriate in the context of very small sample sizes for POM values: an apparent "outlier" opinion in a small cohort may reflect the existence of a genuine faction in a larger population.

We also report a **"pairwise wins"** statistic derived from our sensitivity analysis, roughly indicating the robustness of the ranking to resampling simulations. This was calculated as the percentage of times a given question had higher POM VOI than other questions in the set in a resampling simulation. We use this as an additional robustness check on the main POM results.

Throughout this report, we refer to the probability of the ultimate question resolving positively, "AI causing extinction by 2100", as **P(U)**, and the probability of indicator questions as **P(c)**. **P(U|c)** is the probability of the ultimate question, given that an indicator question

---

[40] Most questions in the main question-rating survey were selected based on high scores from either superforecasters or expert "proxy" judges, or both. However, two questions, EX30 and HS50, were randomly selected from the intersection of the bottom half of superforecaster and expert proxy scores. While these questions ranked poorly among superforecasters in the main survey, EX30 notably received the second-highest score from experts. Overall, the correlation between expert "proxy" scores and expert scores in the main question-rating round was weak.

[41] For a description of Kullback-Leibler VOI, see Appendix 4: VOI technical explanation.

[42] The advantages of POM over straight VOI are (i) it is more interpretable; and (ii) it does not penalize respondents with low prior probability P(U). The size of the update is constrained by the prior probability P(U) together with the probability of the crux event P(c) to be less than P(U) / P(c).

resolves positively. When we report aggregate probabilities, we use the arithmetic mean. We report **relative risk** as P(U|c) / P(U).

How to read a conditional tree diagram

A conditional tree diagram begins with an initial node displaying the “start date”, usually the point in time at which the conditional tree survey was elicited. This node also displays a current estimate of the probability of some “ultimate question,” which may be either an individual’s estimate or an average over respondents.

The subsequent node represents an “indicator,” or an event which implies an update to the probability of the ultimate question. It displays a highly abridged question title and question ID, for which question summaries and full texts can be found in Appendix 1. Below the node is an estimate of the probability of TRUE or FALSE resolution.

The first indicator question may be followed by one or more additional indicator question layers. Resolution of these questions is estimated conditional on the outcomes of any previous question layers. That is, when indicator question #1 resolves positively, it may affect the probability of indicator question #2 resolving positively, and this is reflected in the values displayed in Figure 3.1.

Finally, the ultimate question nodes are the terminal point of each branch, and display an updated probability estimate conditional on the path leading to it.

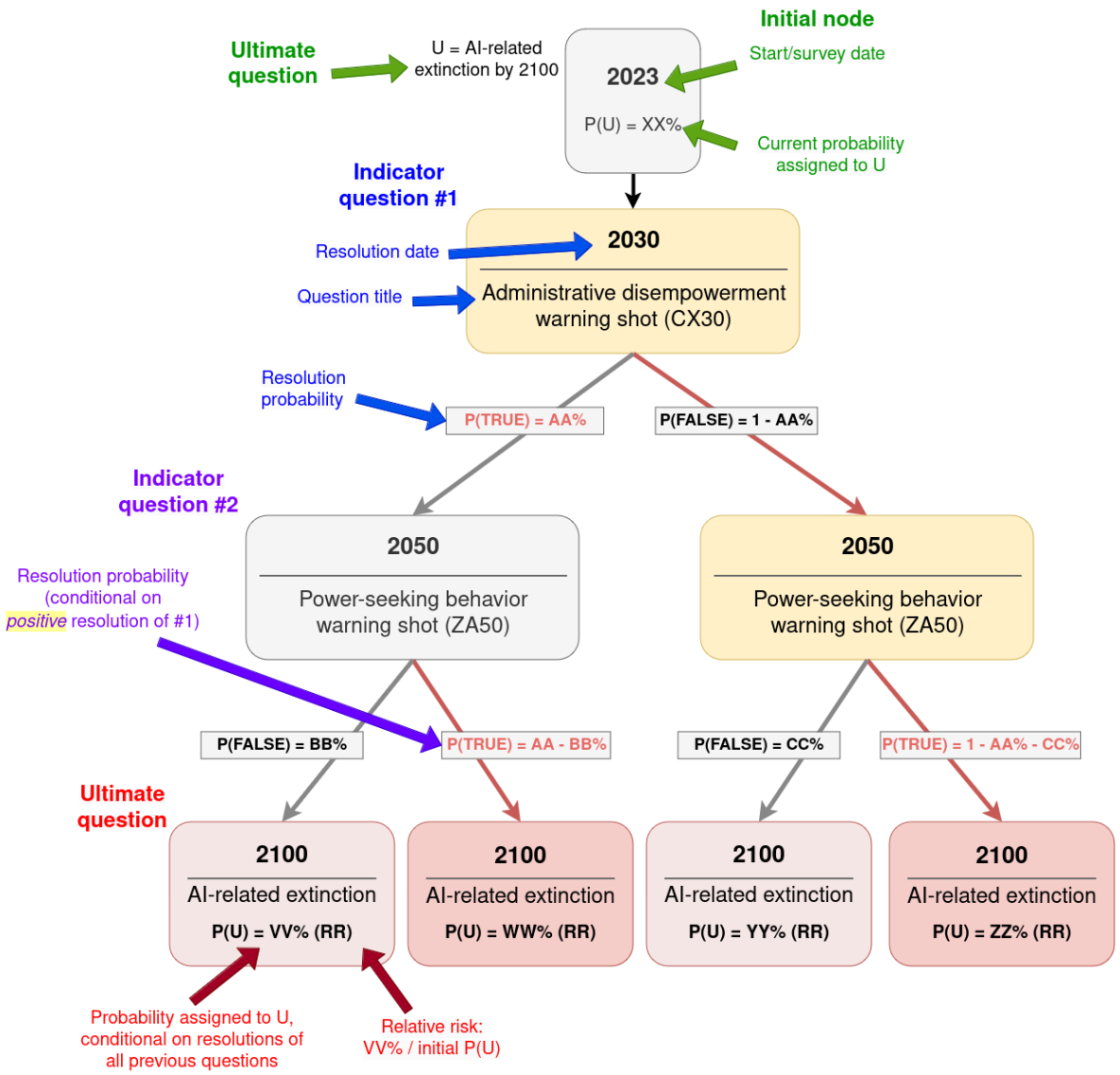


Figure 3.1.1: *Conditional tree diagram for AI-related extinction risk*

3.1 Question ratings summary

Tables 3.1.1 and 3.1.2 show ratings for thirteen questions from the question generation process and two additional, highest-ranked "status quo" questions drawn from forecasting platforms, for a total of fifteen questions. Summaries of question content can be found in Table 3.1.3.

On average, the experts estimated that the probability of AI-related extinction by 2100 is 16.8%. The superforecasters were more skeptical of the risk, with an average probability of 0.25%.[43]

Question rating summary

[43] In the supplementary survey (see Section 4.1), two superforecasters updated their forecasts slightly, resulting in an average P(U) of 0.26%.

	Superforecasters		Experts*	
	VOI rank	**Relative risk** *(P(U|c) / P(U))*	**VOI rank**	**Relative risk** *(P(U|c) / P(U))*
2030 Questions				
Administrative disempowerment warning shot (CX30)	1	13.4	1	1.9
Deep learning revenue (VL30)	2	2.5	4	1.2
Extinction-level pathogens feasible (CQ30)	3	1.9	6	0.8
Deceptive AI warning shot (ZD30)	4	3.2	3	1.1
AI involvement in nuclear arms (HB30)***	5	1.5	NA	NA
Kurzweil/Kapor longbet (STQ9)**	6	1.1	7	0.8
AI arms race, multipolar result (NG30)	7	1.0	5	1.1
AI autonomous purchasing (EX30)	8	1.0	2	1.6
2050-2070 Questions				
AI causing deaths, ineffectual response (CX50)***	1	23.2	NA	NA
Power-seeking behavior warning shot (ZA50)	2	2.4	4	1.4
High AI investment, low safety indicators (VL70)	3	1.3	2	4.2
No aligned AGI (CX70)	4	0.8	1	1.5
AI CEOs / Research productivity (EX50)	5	1.3	5	1.2
Less prosocial behavior / Failing institutions (HS50)	6	1.0	6	0.9
Human-machine intelligence parity (STQ247)**	7	1.0	3	1.4

Table 3.1.1: *Question rating summary*
VOI rank from group POM VOI means. Relative risk is an arithmetic mean of each individual's relative risk (P(U|c) / P(U)).
**Note that each question was shown to a random subset of experts, not to all experts. This may have the effect of amplifying noise due to individual response biases, for both the VOI ranking and relative risk.*
***Denotes external questions not generated as part of the conditional tree process.*
****Denotes questions elicited in a supplementary survey round along with the status quo question set (see section 4.1). This round was only administered to the superforecaster sample.*

Question ratings (all years)

		Superforecasters			Experts		
Question	**Res year**	**Mean POM**	**Mean POM-z**	***n***	**Mean POM**	**Mean POM-z**	***n***
AI causing deaths, ineffectual response (CX50)**	2050	6.34%	0.08	7	NA	NA	NA
Administrative disempowerment warning shot (CX30)	2030	3.55%	0.13	8	1.26%	0.94	5
Deep learning revenue (VL30)	2030	1.68%	-0.04	7	0.64%	0.16	5
Power-seeking behavior warning shot (ZA50)	2050	1.59%	0.53	8	3.00%	0.56	5
Extinction-level pathogens feasible (CQ30)	2030	1.37%	0.57	8	0.18%	-0.59	5

Deceptive AI warning shot (ZD30)	2030	0.98%	0.23	8	0.85%	0.10	5
AI involvement in nuclear arms (HB30)**	2030	0.68%	-0.07	7	NA	NA	NA
High AI investment, low safety indicators (VL70)	2070	0.54%	0.67	8	10.19%	-0.05	5
No aligned AGI (CX70)	2070	0.37%	-0.21	8	14.71%	0.53	6
Kurzweil/Kapor longbet (STQ9)*	2030	0.27%	0	8	0.15%	-0.41	5
AI CEOs / Research productivity (EX50)	2050	0.26%	-0.17	8	1.12%	-0.59	4
Less prosocial behavior / Failing institutions (HS50)	2050	0.26%	-0.30	8	0.25%	-0.63	6
AI arms race, multipolar result (NG30)	2030	0.26%	-0.28	8	0.37%	-0.33	4
Human-machine intelligence parity (STQ247)*	2040	0.14%	-0.59	8	4.19%	0.11	4
AI autonomous purchasing (EX30)	2030	0.02%	-0.55	8	0.98%	0.06	4

Table 3.1.2: *Question ratings (all years)*
**Denotes external questions not generated as part of the conditional tree process.*
***Denotes questions elicited in a supplementary survey round along with the status quo question set (see section 4.1). This round was only administered to the superforecaster sample.*

All main question-rating survey questions

Year	Title	Concise question summary
2030	Administrative disempowerment warning shot (CX30)	AI systems are preventing a substantial number of people from accessing financial assets, voting, moving freely, accessing the internet, or accessing certain types of housing or employment.
2030	Deep learning revenue (VL30)	Revenue from deep learning doubles every two years before 2030.
2030	Extinction-level pathogens feasible (CQ30)	A panel of relevant experts believes extinction-level pathogens are both possible and will be practically feasible relatively soon, at relatively low cost.
2030	Deceptive AI warning shot (ZD30)	After a costly incident, AI developers or relevant experts conclude that a top lab's AI system behaved deceptively.
2030	AI involvement in nuclear arms (HB30)**	Governments publicly integrate AI decision-making into their nuclear weapons control systems, or nuclear risk experts worry about AI.
2030	Kurzweil/Kapor longbet (STQ9)*	Before 2030, will an AI complete the Turing Test in the Kurzweil/Kapor Longbet?
2030	AI arms race, multipolar result (NG30)	AI-related government spending is high for both the US and China, with low research collaboration, and relative economic parity; OR private investment in AI is very high and there is no single dominant company.
2050	AI causing deaths, ineffectual response (CX50)**	AI systems cause the deaths of >1m humans, which is not intended by any humans, and subsequently a panel of experts believes the collective global response has not noticeably reduced risk of similar events.
2050	Power-seeking behavior warning shot (ZA50)	AI developers shut down an expensive AI system after it displays a power-seeking behavior, such as hoarding resources, interfering with vital infrastructure, propagating itself, etc.
2070	High AI investment, low safety indicators (VL70)	Compute spending is high and experts agree that aligning AI systems is very difficult; and there is insufficient political attention to

		AI safety.
2070	No aligned AGI (CX70)	No AI system exists which both performs well on general ability benchmarks (e.g. Q&A dataset) and has positive indicators of alignment (performance on alignment benchmarks, confidence of AI safety researchers).
2050	AI CEOs / Research productivity (EX50)	AI systems are performing entire roles at top companies that currently are performed by C-suite executives; or research productivity is higher than it was in 1930.
2050	Less prosocial behavior / Failing institutions (HS50)	Charitable donations in the US have fallen dramatically; or corruption rises dramatically in the US or Europe; or autocracy increases dramatically worldwide.
2040	Human-machine intelligence parity (STQ247)*	Will there be Human-machine intelligence parity before 2040?
2030	AI autonomous purchasing (EX30)	AI autonomously buying goods or services (e.g. purchasing flights, managing inventories for companies, etc) -- >$1 million / yr

Table 3.1.3: *All main question-rating survey questions*
Question IDs link to the full text of the question operationalization in Appendix 1.
**Denotes external questions not generated as part of the conditional tree process.*
***Denotes questions elicited in a supplementary survey round along with the status quo question set (see section 4.1). This round was only administered to the superforecaster sample.*

3.2 Candidate high VOI trees from two camps

This section displays high VOI trees produced by the main question-rating survey data for skeptical superforecasters and for concerned experts. For each group, we included a selection of the most informative questions in the tree. Only the superforecaster tree is a true conditional tree, as only superforecasters were surveyed on every combination of the top-scoring questions.

Skeptical superforecasters' conditional tree

We surveyed the superforecasters in our sample for conditional forecasts on sixteen scenarios. These scenarios were combinations of the top-ranked questions: "administrative disempowerment" (CX30), "extinction-level pathogens" (CQ30), "AI-related deaths" (CX50) and "Power-seeking" (ZA50).[44] Seven superforecasters responded. The sixteen scenarios are mutually exclusive and exhaust the space of possible outcomes; thus, we ensured that each respondent's probabilities assigned to the scenarios summed to 100% and showed them their *implied P(U)*, the average of their P(U|scenario)'s weighted by the likelihood they assigned to each scenario (see Figure 2.2.2). We averaged the forecasts for each P(scenario) and P(U|scenario) separately to create an aggregate judgment. The implied

[44] The goal was to choose the most informative questions. The initial selection criteria were to choose the top-ranked question by POM and POM-z for questions resolving in 2030 and 2050-2070 separately, including both where these disagreed. For 2030, we chose CX30 (highest POM) and CQ30 (highest POM-z). For 2050-2070, we chose CX50 based on it having the highest POM. While the selection criteria suggested that VL70 should be selected as the top POM-z question, as a whole the evidence pointed to ZA50 being more informative (higher POM, at 1.59% vs 0.54%; POM-z close to VL70, at 0.53 vs 0.67; and higher under the pairwise wins robustness check, at 87% vs 64%).

P(U) of this aggregate was then used to compute average relative risk (the multiplier in each branch of the tree). A simplified version of the resulting tree is shown in Figure 3.2.1.

For example, conditional on both “Extinction-level pathogens” and “AI-related deaths” resolving positively (superforecasters assign a 2.82% chance to this outcome), the superforecasters would on average update their P(U) from 0.94% to 6.21%.

The scenario that would constitute the biggest update is the case where all four questions that would imply higher risk resolve positively. If the four relevant risk-increasing outcomes were to happen (far right in the full tree (a)), the superforecasters’ relative risk assessment is 10.7 (i.e., they would be 10.7x more concerned than they currently are about the risk of AI-related extinction). Conversely, if none of the questions resolve positively (far left), their relative risk assessment is 0.3.

Note that the average P(U) in this survey (0.94% in Figure 3.2.1) is higher than in the main survey (0.25%), which we used to compute VOI. Two superforecasters made substantial updates to their unconditional probability of AI-related extinction by 2100 (P(U)) between the main survey (conducted in July 2023) and this combinations survey (conducted in February to March 2024 with a follow-up in May), which may be attributable to events of the intervening months or to the exercise of thinking through scenarios. One superforecaster updated from 0.1% to 0.4% and another from 1% to 4.2%. The other five did not update.

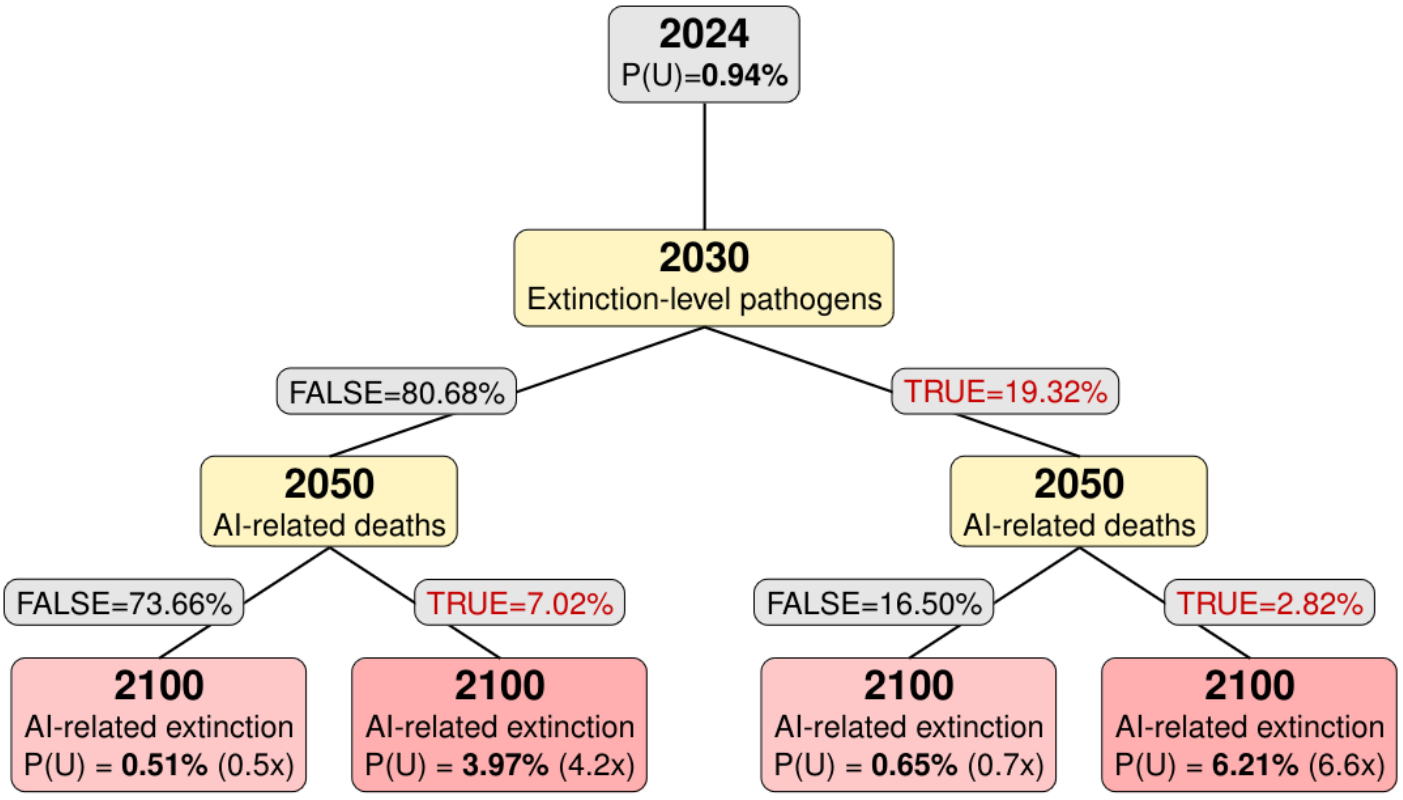


Figure 3.2.1: *Skeptical superforecaster conditional tree*
This is a collapsed tree of combinations of the superforecasters’ highest-VOI questions. For the purpose of legibility, we are presenting a simplified tree, using two of the four questions. We collapsed the sixteen scenarios into four combinations. Positive resolution (“TRUE”) is a bad outcome for both questions. The far right scenario (both TRUE) constitutes the worst scenario, a 6.6x update, and the far left scenario is the best (both FALSE) with a halving of the superforecasters' current risk estimate. You can see the full, unpruned tree here (a).

Concerned experts' conditional trees

Figure 3.2.2 presents the question from each year (2030, 2050, and 2070) that surveyed experts rated the highest, on average, in terms of POM VOI. As a whole, among these highest-POM VOI questions, the experts would be most worried if there were an administrative disempowerment warning shot by 2030 (1.9x update from their current unconditional P(U) of 17%). Conversely, if we do not see a power-seeking behavior warning shot by 2050, the experts would be least worried (0.6x update).

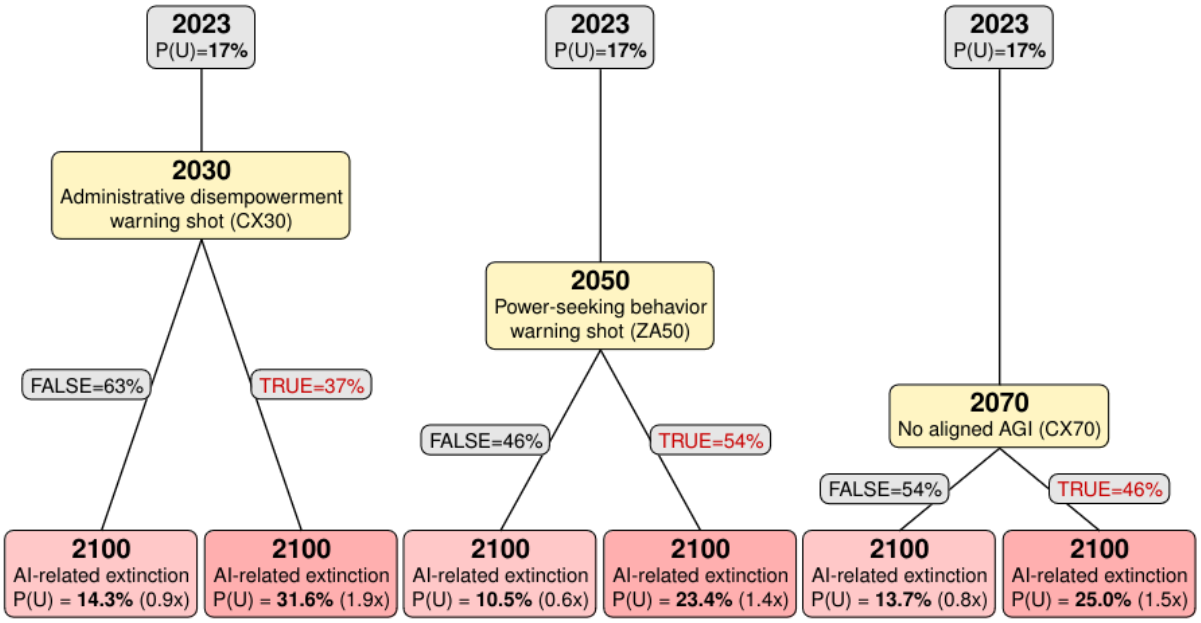


Figure 3.2.2: *A diagram showing how experts update on three questions for different resolution years that scored particularly well on our VOI metric. Since experts answered different sets of questions, we derived P(U|C) and P(U|~C) (the probabilities on the bottom level) by multiplying the whole expert group's average P(U) of 17% by the average relative risk factor for each crux.*[45]

[45] Careful readers will note that the probabilities in this figure do not yield the mean POM VOI values we report (see Tables 3.4.1 and 3.4.2). Mean POM VOI tells us how valuable a crux is for a group, on average, by computing POM VOI at the individual level and then aggregating. The average relative updates, across individuals in the same group, sometimes tells a quite different story.

3.3 Skeptical superforecasters' question ratings

2030 questions

Question	Mean POM	P(c)	RR *(P(U|c) / P(U))*	Mean POM-z	Pairwise wins	*n*
Administrative disempowerment warning shot (CX30)	3.55%	16%	13	0.13	83%	8
Deep learning revenue (VL30)	1.68%	33%	2.5	-0.04	59%	7
Extinction-level pathogens feasible (CQ30)	1.37%	39%	1.9	0.57	75%	8
Deceptive AI warning shot (ZD30)	0.98%	32%	3.2	0.23	64%	8
AI involvement in nuclear arms (HB30)**	0.68%	18%	1.5	-0.07	50%	7
Kurzweil/Kapor longbet (STQ9)*	0.27%	43%	1.1	0	33%	8
AI arms race, multipolar result (NG30)	0.26%	39%	1.0	-0.28	33%	8
AI autonomous purchasing (EX30)	0.02%	35%	1.0	-0.55	3%	8

Table 3.3.1: *Skeptical superforecasters' 2030 question ratings*
P(c) is the arithmetic mean of this group's responses. RR (relative risk) is an arithmetic mean of each individual's relative risk (P(U|c) / P(U)).
**Denotes external questions not generated as part of the conditional tree process.*
***Denotes questions elicited in a supplementary survey round along with the status quo question set (see section 4.1). This round was only administered to the superforecaster sample.*

Skeptical superforecasters' top-rated question by mean POM was "Administrative disempowerment warning shot" (CX30), referring to a scenario in which *"AI systems are preventing a substantial number of people from accessing financial assets, voting, moving freely, accessing the internet, or accessing certain types of housing or employment."* It scored ~3.6% of the theoretical maximum VOI score on average. However, this high value was driven by a single respondent, with the question achieving a remarkable 25% of the theoretical maximum VOI for this individual.[46] This is consistent with superforecasters in our sample preferring questions which refer to concrete AI-related harms, though the high variance in VOI ratings for this question suggest that there is no consensus on exactly *which* harms provide the clearest signal.

The top-rated question by POM-z, "Feasibility of extinction-level pathogens" (CQ30), refers to a scenario in which *"A panel of relevant experts believes extinction-level pathogens are both possible and will be practically feasible relatively soon, at relatively low cost."* It is the question that respondents most agreed was informative, though the highest VOI rating any individual gave this question was only 5.2% of the theoretical maximum. Interestingly, this

[46] While an extreme data point could typically indicate a coding error, the subcomponents of VOI analysis suggest a genuine answer rather than a common error such as a misplaced decimal. The outlier respondent assigned a low probability (0.5%) to the "administrative disempowerment warning shot" scenario, but provided a substantial update (a 100-fold increase, from 0.1% to 10%) toward AI extinction if the scenario were to occur. In contrast, all other respondents thought the probability of it occurring was higher (mean=18%), but offered smaller updates than the outlying respondent (mean = 1x, with three updating not at all and one updating down).

question does not refer to realized harm, but rather to favorable conditions for harm to take place. Such questions may gain a VOI advantage by omitting divisive or low-probability conditions that hinge on human motivations for misusing AI technologies.[47] It was the third most likely 2030 question to resolve positively.

No mean POM differences between questions were significant in this sample (after correcting for multiple testing using the Bonferroni correction, all p-values were equal to 1). Survey responses between filtering and main survey rounds were fairly similar, though with some notable differences. See Appendix 2.1 for further details on intra-individual response variability.

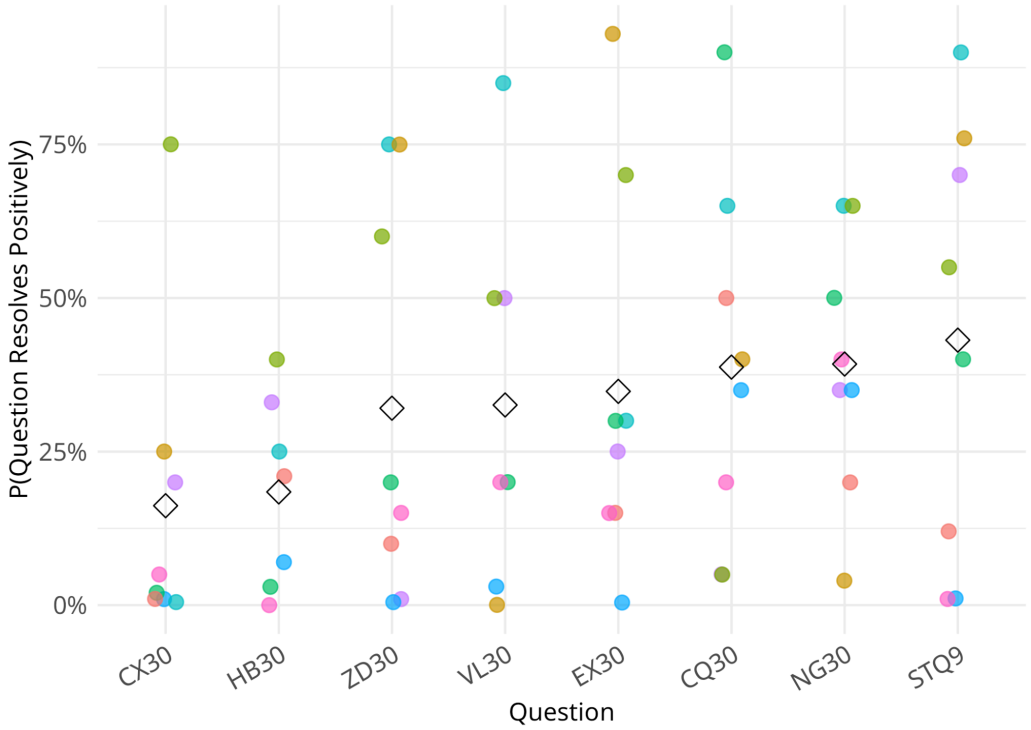


Figure 3.3.1: Skeptical superforecasters' 2030 P(c)

[47] Or, indeed, the motivations of a misaligned AI system with access to weaponizable technology.

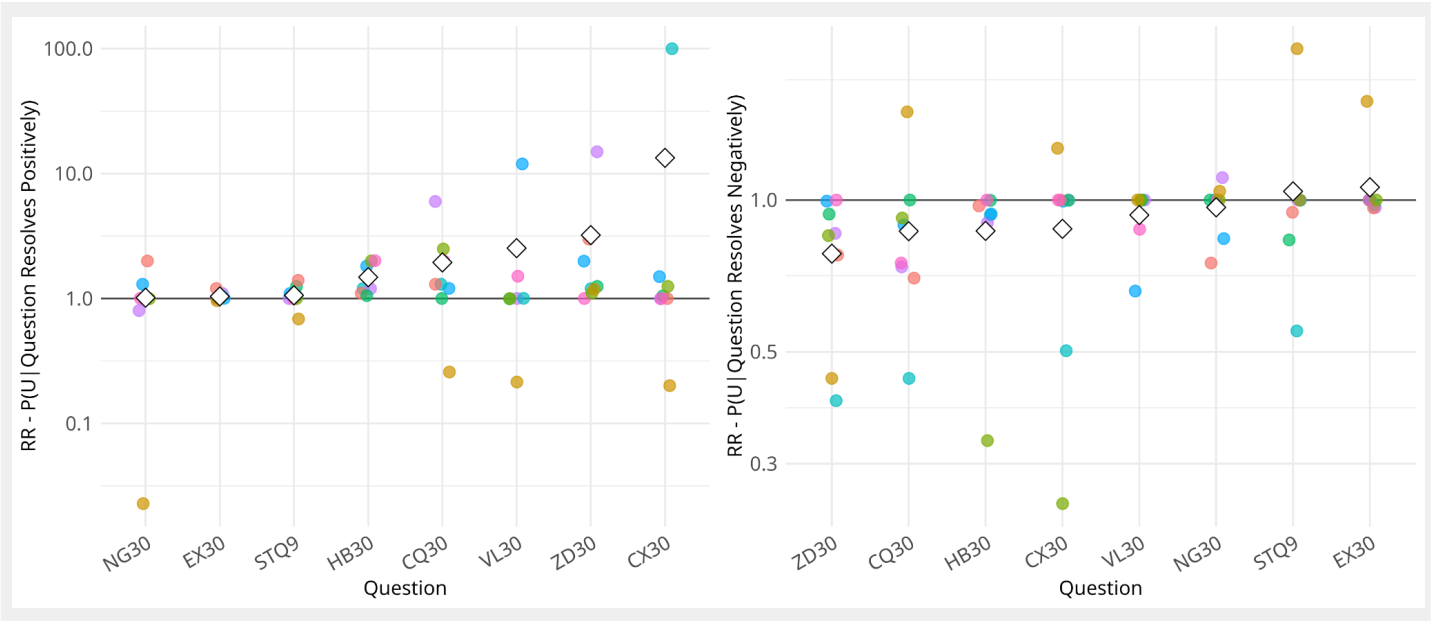


Figure 3.3.2: *Skeptical superforecasters' 2030 relative risk. Diamonds represent arithmetic means. Log scale. Relative risk >1 reflects a positive update, that is, where P(U|c) > P(U).*

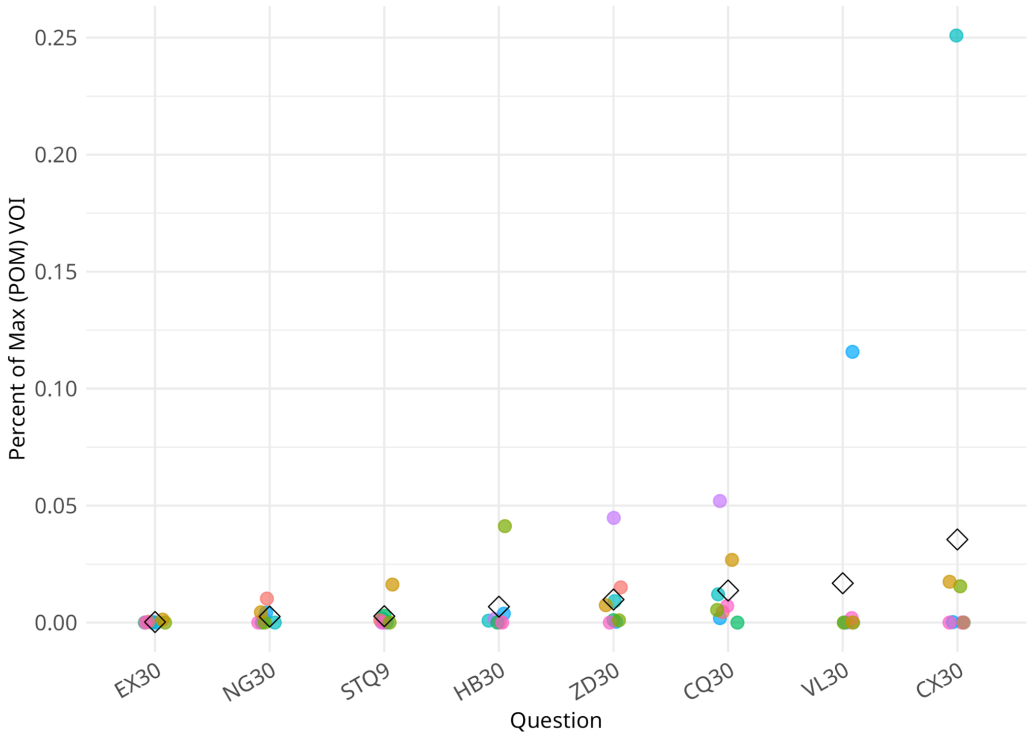


Figure 3.3.3: *Skeptical superforecasters' 2030 POM VOI. Diamonds represent arithmetic means.*

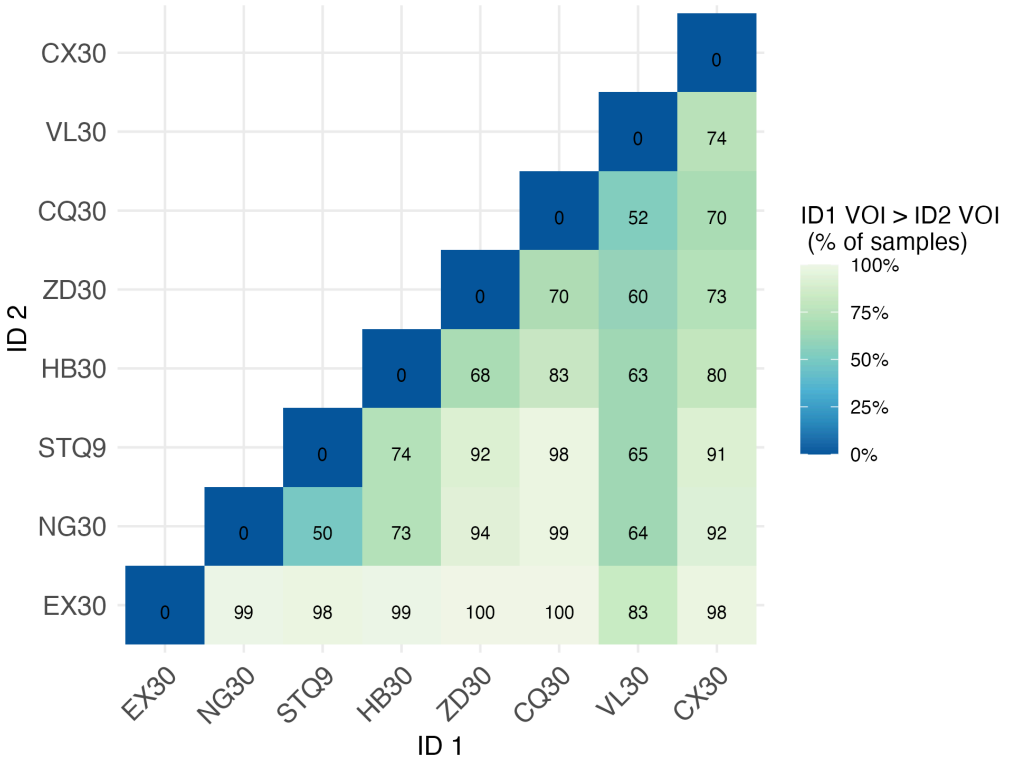


Figure 3.3.4: *Skeptical superforecasters' 2030 POM VOI sensitivity matrix (pairwise wins). Visualization of resampling simulation results.*

2050-2070 questions

Question	Mean POM	P(c)	RR *(P(U|c) / P(U))*	Mean POM-z	Pairwise wins	*n*
AI causing deaths, ineffectual response (CX50)**	6.34%	6%	23	0.08	67%	7
Power-seeking behavior warning shot (ZA50)	1.59%	38%	2.4	0.53	87%	8
High AI investment, low safety indicators (VL70)	0.54%	38%	1.3	0.67	64%	8
No aligned AGI (CX70)	0.37%	34%	0.8	-0.21	48%	8
AI CEOs / Research productivity (EX50)	0.26%	21%	1.3	-0.17	35%	8
Less prosocial behavior / Failing institutions (HS50)	0.26%	31%	1.0	-0.30	32%	8
Human-machine intelligence parity (STQ247)*	0.14%	53%	1.0	-0.59	17%	8

Table 3.3.2: *Skeptical superforecasters' 2050-2070 question ratings. P(c) is the geometric mean of odds of this group's responses. RR (relative risk) is an arithmetic mean of each individual's relative risk (P(U|c) / P(U)).*
**Denotes external questions not generated as part of the conditional tree process.*
***Denotes questions elicited in a supplementary survey round along with the status quo question set (see section 4.1). This round was only administered to the superforecaster sample.*

Skeptical superforecasters' top-rated question by mean POM was "AI causing deaths, ineffectual response" (CX50), referring to a scenario in which *"AI systems cause the deaths of >1m humans, which is not intended by any humans, and subsequently a panel of experts believes the collective global response has not noticeably reduced risk of similar events."* It scored ~6.3% of the theoretical maximum VOI score on average. Similarly to the top-scoring 2030 question, this question benefits from a single strong opinion, and thus does relatively worse on both sensitivity analysis and POM-z. It is an even starker example of concrete harm caused by AI systems; but was also among the least likely questions to resolve positively, at 6%.

While neither first-ranked by mean POM or POM-z, the most robust question in sensitivity analysis was "Power-seeking behavior warning shot" (ZA50), in which *"AI developers shut down an expensive AI system after it displays a power-seeking behavior, such as hoarding resources, interfering with vital infrastructure, propagating itself, etc."* This question does not fit as clearly with superforecasters' apparent preference for questions referring to concrete harm.

The highest rated question by POM-z was "High AI investment, low safety indicators" (VL70), in which "*Compute spending is high and experts agree that aligning AI systems is very difficult; and there is insufficient political attention to AI safety."* However, in absolute terms it was rated relatively low at 0.5% POM.

"No aligned AGI" (CX70) is unique in this question set as the only question which on average updated superforecasters *away* from AI-related extinction (mean relative risk = 0.8x). Here respondents may have inferred that a world with no aligned AGI by 2070 was more likely to be a world with no AGI of any kind, than a world with only *unaligned* AGI.

No mean POM differences between questions were significant in this sample (after correcting for multiple testing, all p-values were equal to 1). Survey responses between

filtering and main survey rounds were broadly similar. See Appendix 2.1 for further details on intra-individual response variability.

2030 vs 2050/2070 questions

In the AICT question set (that is, all questions excluding STQ9 and STQ247) the average of POM responses for 2030 questions, 1.2%, was slightly lower than that of 2050-2070 questions, at 1.5%. Due to the small number of questions included and the small absolute size of the difference, this does not seem indicative of a genuine VOI difference between earlier and later questions in our set.

Responses for 2030 and 2050-2070 AICT questions suggested similar probabilities of positive resolution, 30% and 28% respectively (mean; interquartile range (IQR) = 5% - 50% and 10% - 40%).[48] However, the relative risk was lower for 2030 questions at 3.6x (mean; IQR = 1x - 1.5x), vs. 2050 - 2070 questions at 4.6x (mean; IQR = 1x - 1.5x).

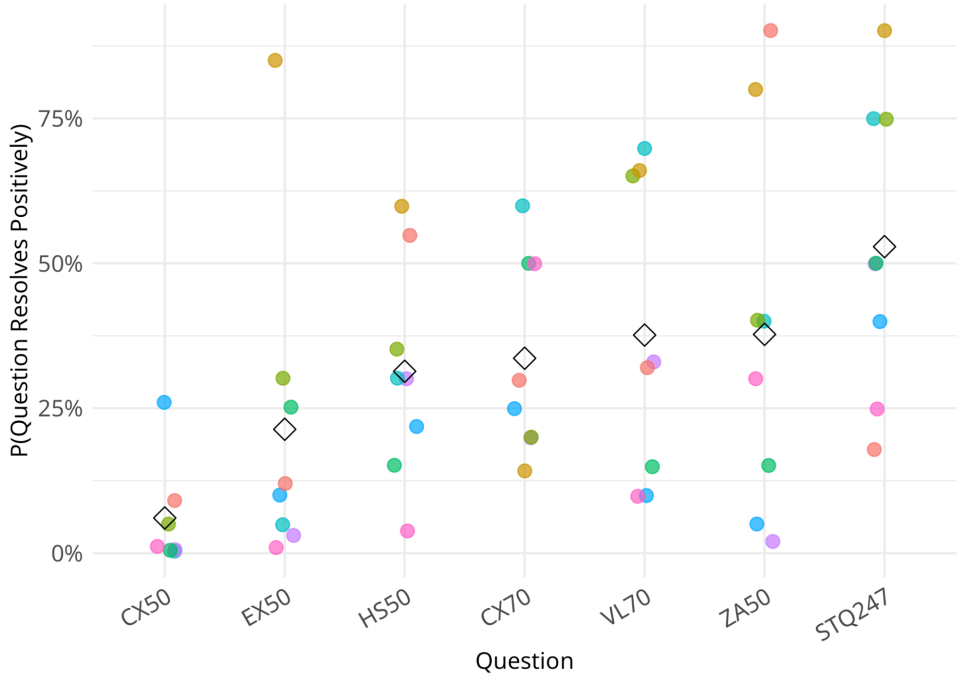


Figure 3.3.5: *Skeptical superforecasters' 2050-2070 P(c)*

[48] Interquartile range (IQR) is the middle 50%, or the difference between the 25th and 75th percentile forecasts.

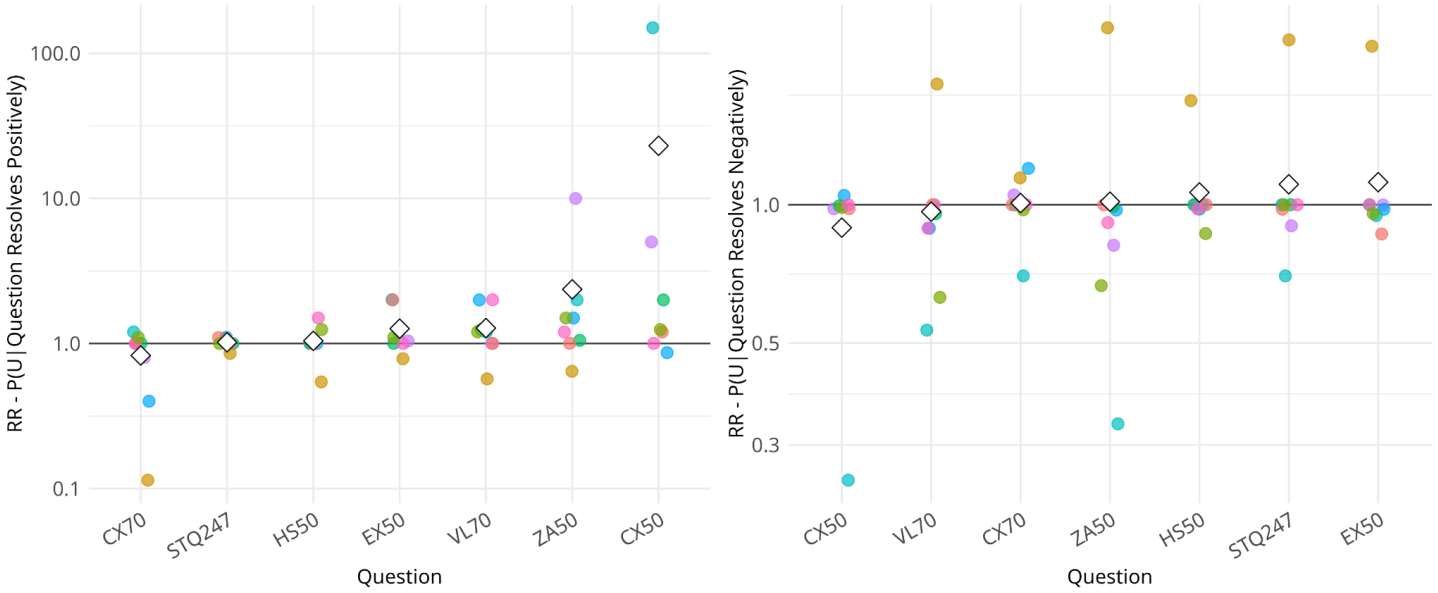


Figure 3.3.6: *Skeptical superforecasters' 2050-2070 relative risk. Diamonds represent mean values. Log scale. Relative risk >1 reflects a positive update, that is, where P(U|c) > P(U).*

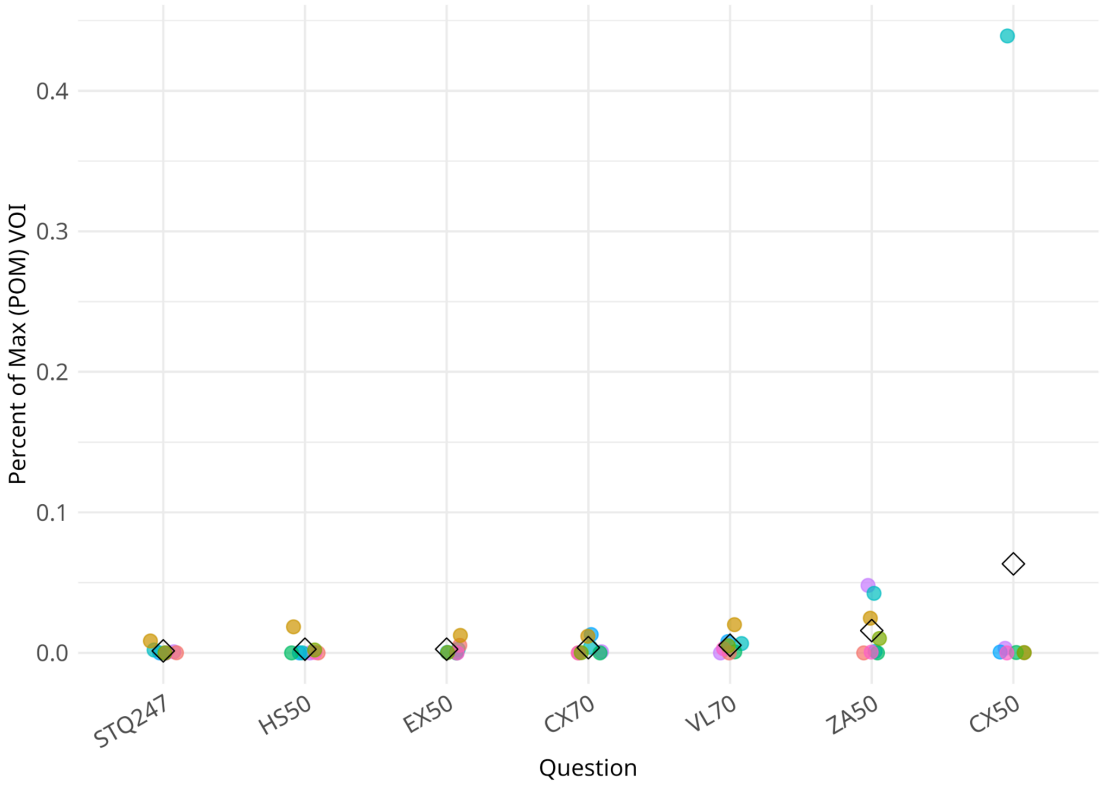


Figure 3.3.7: *Skeptical superforecasters' 2050-2070 POM VOI. Diamonds represent arithmetic means.*

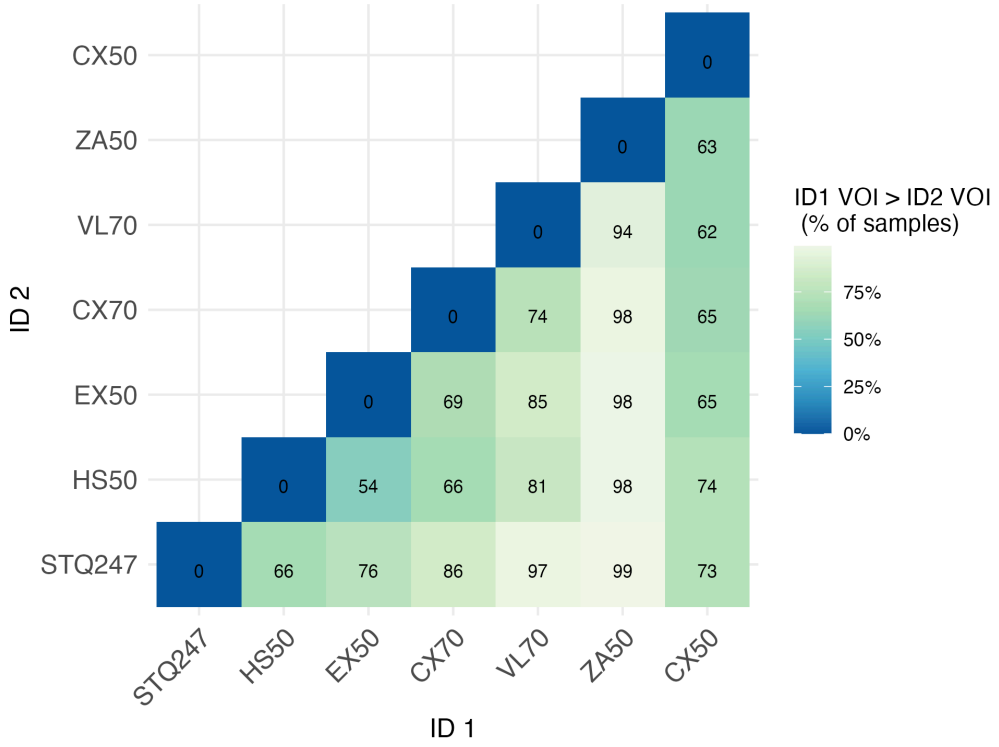


Figure 3.3.8: *Skeptical superforecasters' 2050/2070 POM VOI sensitivity matrix (pairwise wins). Visualization of resampling simulation results.*

3.4 Concerned experts' question ratings

2030 questions

Question	Mean POM	P(c)	RR *(P(U|c) / P(U))*	Mean POM-z	Pairwise wins	*n*
Administrative disempowerment warning shot (CX30)	1.26%	37%	1.9	0.94	87%	5
AI autonomous purchasing (EX30)	0.98%	54%	1.6	0.06	75%	4
Deceptive AI warning shot (ZD30)	0.85%	66%	1.1	0.10	66%	5
Deep learning revenue (VL30)	0.64%	17%	1.2	0.16	48%	5
AI arms race, multipolar result (NG30)	0.37%	38%	1.1	-0.33	41%	4
Extinction-level pathogens feasible (CQ30)	0.18%	25%	0.8	-0.59	18%	5
Kurzweil/Kapor longbet (STQ9)*	0.15%	57%	0.8	-0.41	15%	5

Table 3.4.1: *Concerned experts' 2030 question ratings. P(c) is the arithmetic mean of this group's responses. RR (relative risk) is an arithmetic mean of each individual's relative risk (P(U|c) / P(U)).*
**Denotes external questions not generated as part of the conditional tree process.*

Concerned experts' top-rated question, "Administrative disempowerment warning shot" (CX30), refers to a scenario in which *"AI systems are preventing a substantial number of people from accessing financial assets, voting, moving freely, accessing the internet, or accessing certain types of housing or employment."* It scores best on both POM and POM-z; and while its position is bolstered by an outlier, it is also generally rated well among respondents.[49]

No mean POM differences between questions were significant in this sample (after correcting for multiple testing, all p-values were equal to 1). The filtering round elicitation for these questions appeared to be a poor proxy for expert judgments in the main survey round (see the "Methods" section for more details on the filtering round elicitation).[50]

[49] The outlier respondent assigns a low probability to the question (5%), but updates substantially (relative risk = 3x), while on average respondents rated the question as having moderate probability (mean=37%) and a moderate relative risk (mean=1.9x).
[50] Proxy ratings for 2030 questions showed strong negative correlation with POM VOI judgments from the small sample of experts in the main survey. They also showed slight negative correlation with the main survey POM-z. Notably, a question randomly chosen from the bottom half of proxy scores ranked second by expert POM (EX30). This suggests that many questions from our larger 2030 set might have performed better than the average question in our main question-rating survey if presented to these particular experts.

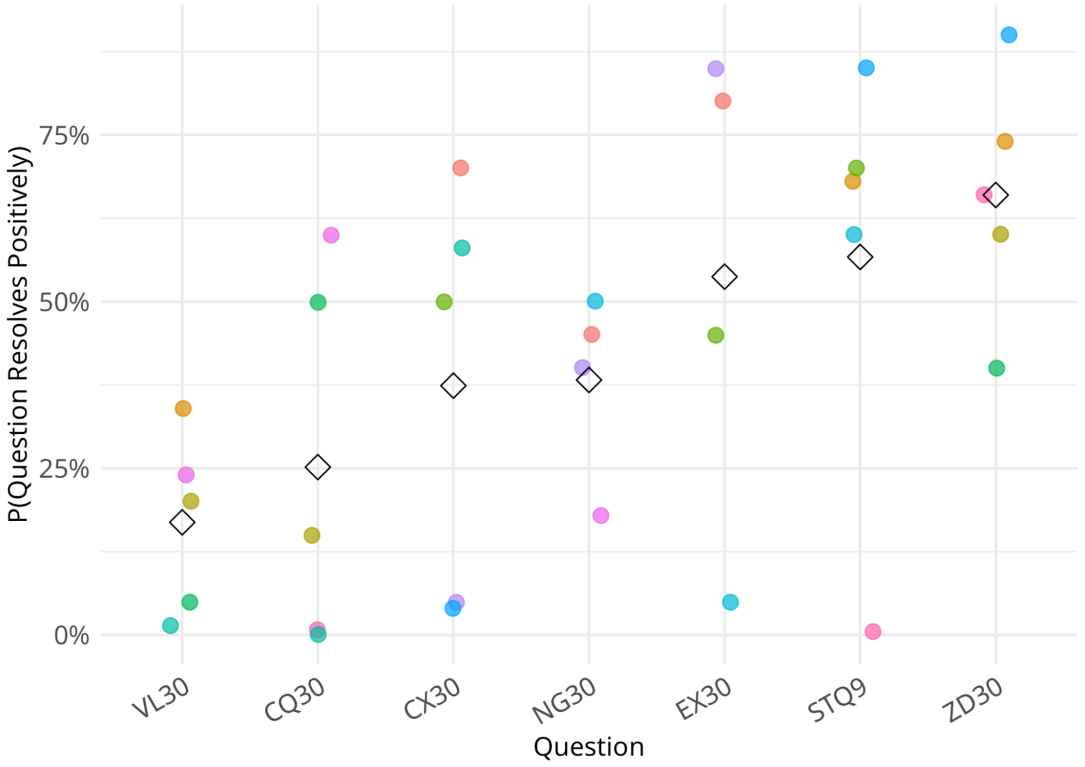


Figure 3.4.1: *Concerned experts' 2030 P(c)*

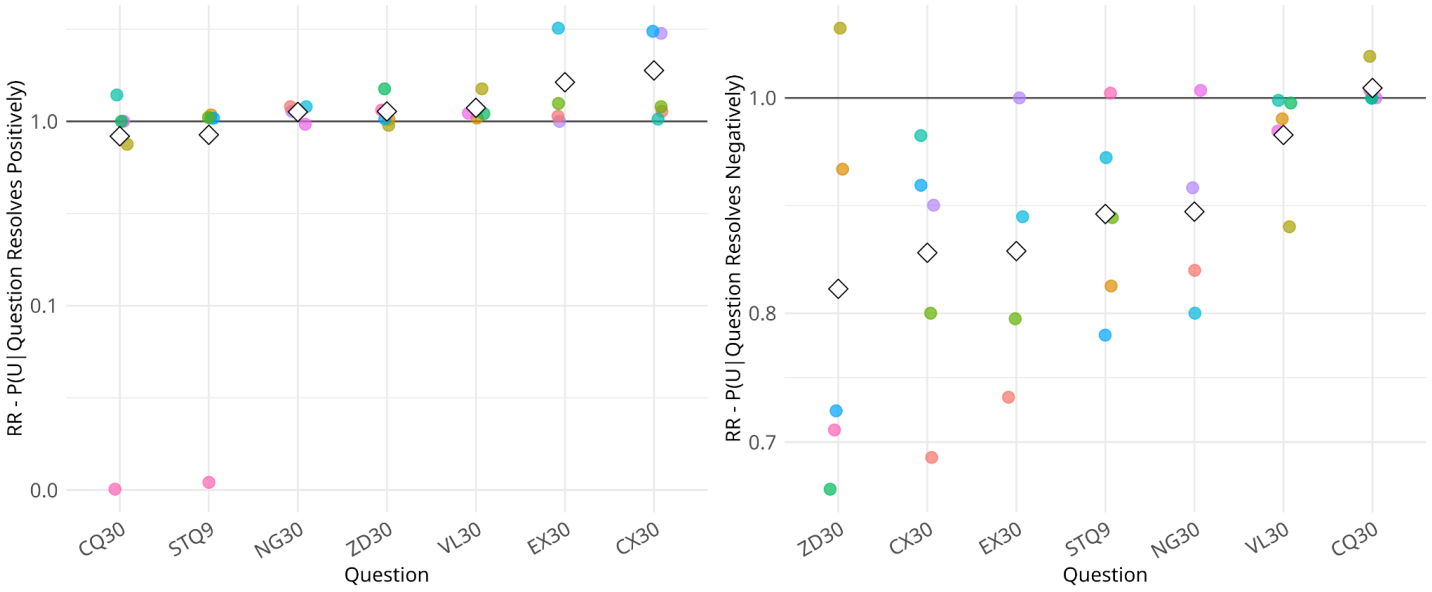


Figure 3.4.2: *Concerned experts' 2030 relative risk. Diamonds represent mean values. Log scale. Relative risk >1 reflects a positive update, that is, where P(U|c) > P(U).*

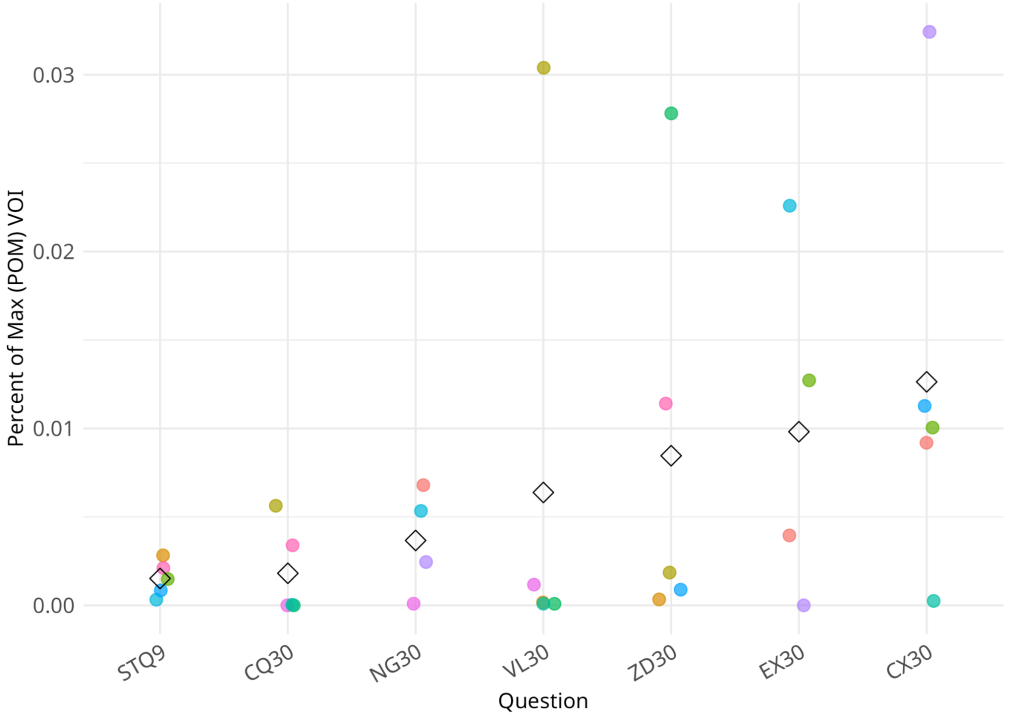


Figure 3.4.3: *Concerned experts' 2030 POM VOI. Diamonds represent arithmetic means.*

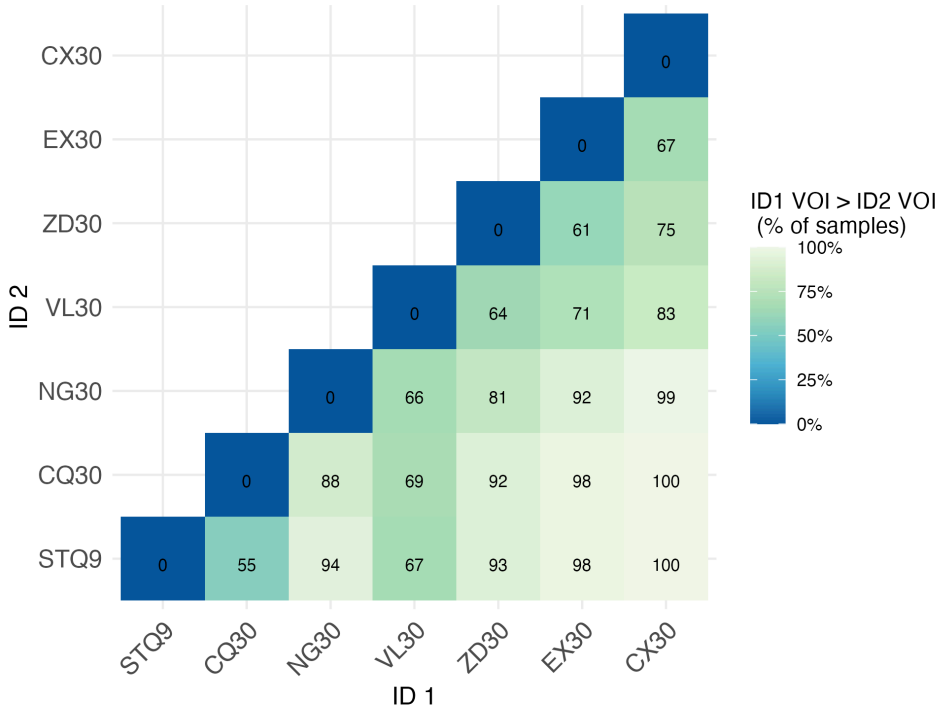


Figure 3.4.4: *Concerned experts' 2030 POM VOI sensitivity matrix (pairwise wins). Visualization of resampling simulation results.*

2050-2070 questions

Question	Mean POM	P(c)	RR *(P(U|c) / P(U))*	Mean POM-z	Pairwise wins	*n*
No aligned AGI (CX70)	14.71%	46%	1.5	0.53	95%	6

High AI investment, low safety indicators (VL70)	10.19%	19%	4.2	-0.05	80%	5
Human-machine intelligence parity (STQ247)*	4.19%	60%	1.4	0.11	56%	4
Power-seeking behavior warning shot (ZA50)	3.00%	54%	1.4	0.56	47%	5
AI CEOs / Research productivity (EX50)	1.12%	46%	1.2	-0.59	22%	4
Less prosocial behavior / Failing institutions (HS50)	0.25%	43%	0.9	-0.63	0%	6

Table 3.4.2: *Concerned experts' 2050-2070 question ratings. P(c) is the arithmetic mean of this group's responses. RR (relative risk) is an arithmetic mean of each individual's relative risk (P(U|c) / P(U)).*
**Denotes external questions not generated as part of the conditional tree process.*

Concerned experts' top-rated question by POM was "No aligned AGI" (CX70), which not only ranked well among this set, but also achieved a very high absolute percentage of maximum VOI of nearly 15%. This question also performed very well on sensitivity analysis, and was judged to be highly probable for this question set at 45.86%. It carried the second highest relative risk at 1.5x, but no respondents gave extremely high relative risk estimates.

The top question by POM-z, "Power-seeking behavior warning shot" (ZA50), had only middling rank by POM, but nonetheless an objectively high POM value of 3%. It was judged to be highly probable at 53.6%, with a moderate relative risk (mean=1.4x).

No mean POM differences between questions were significant in this sample (after correcting for multiple testing, the closest to significance was CX70 vs. HS50 at p = 0.638). The filtering round elicitation for these questions appeared to be a moderately good proxy for expert judgments in the main survey round (see "Methods" section for more details on the filtering round elicitation).[51]

2030 vs 2050/2070 questions

Overall, this set of experts seems to have judged the 2050/2070 set of questions as more informative than the 2030 set: they on average achieved a POM of 5.9%, vs. 2030 questions at 0.63% (2030 IQR = 0.02% - 0.92%; 2050/2070 IQR = 0.18% - 6.6%). This difference appears to be a genuine result, with $p = .043$; it is robust to the removal of any particular question or respondent.

Probability of positive resolution looks quite similar between 2030 and 2050-2070 questions, at 42% and 44% respectively (2030 IQR = 15% - 66%; 2050-2070 IQR = 30% - 60%). Relative risk for later questions was higher in our sample, with an average of 1.8x vs. 2030 questions at 1.2x (2030 IQR = 1.0 - 1.2x; 2050-2070 IQR = 1.0 - 2.0x).

[51] The 2050/2070 proxy performed moderately well for our small expert sample, with a correlation between mean expert POM and proxy rank of -0.4, and mean expert POM-z score and proxy rank of -0.5 (a more negative value indicates a stronger correlation, as higher rank orders are considered worse, while higher VOI scores are better).

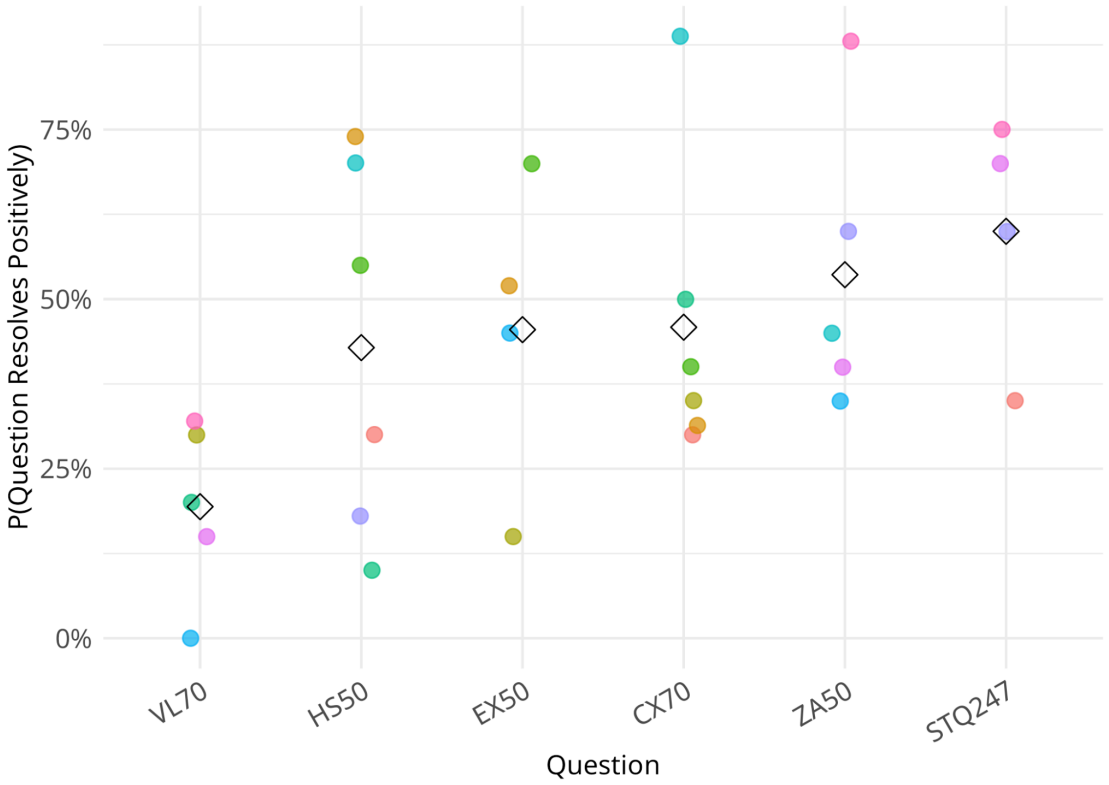


Figure 3.4.5: *Concerned experts' 2050-2070 P(c)*

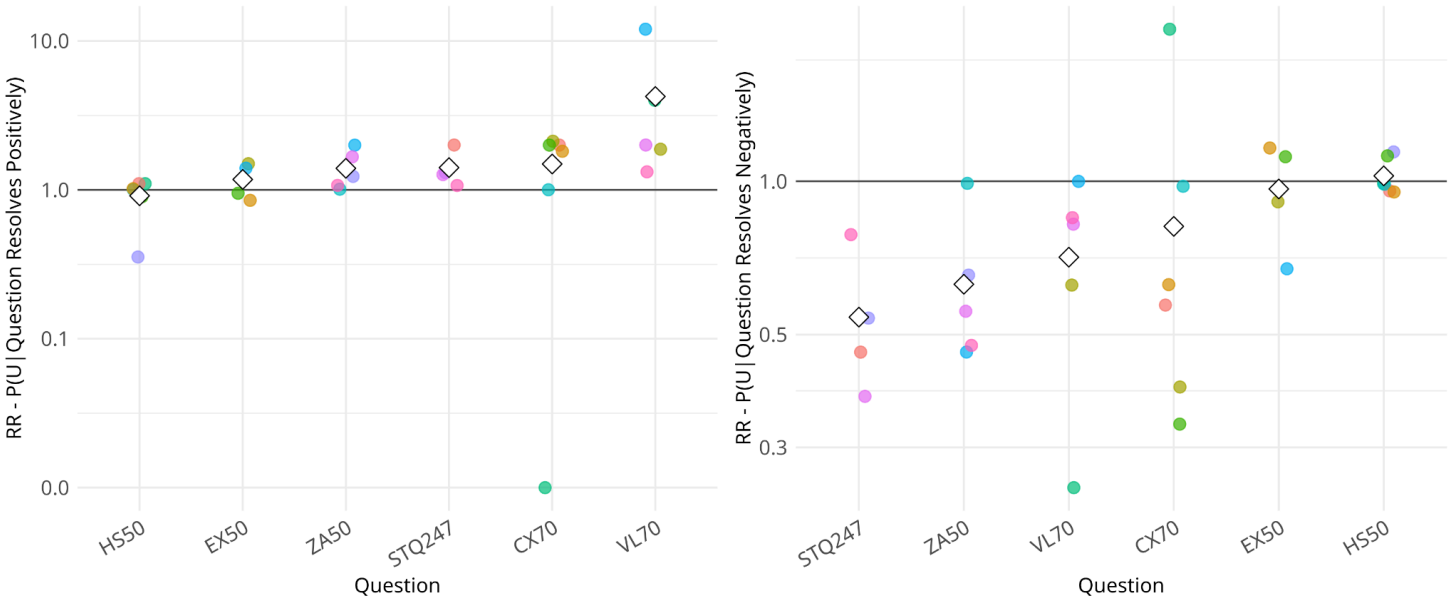


Figure 3.4.6: *Concerned experts' 2050-70 relative risk. Diamonds represent arithmetic means. Log scale. Relative risk >1 reflects a positive update, that is, where P(U|c) > P(U).*

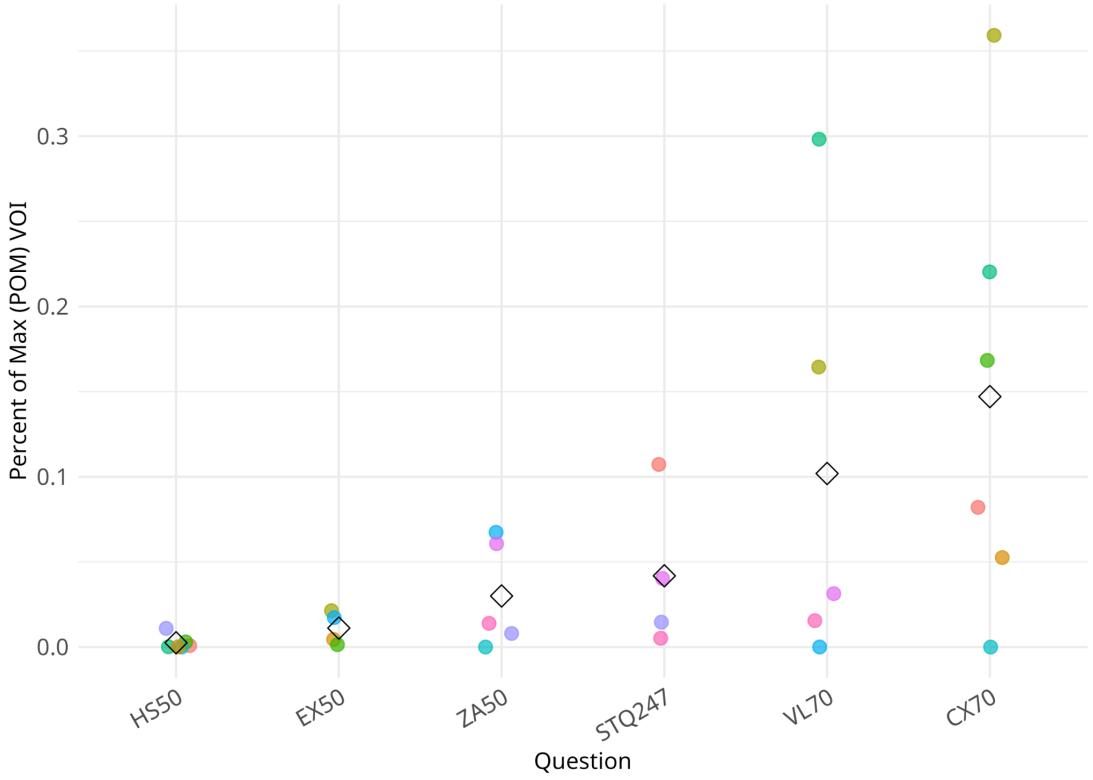


Figure 3.4.7: *Concerned experts' 2050-2070 POM VOI. Diamonds represent arithmetic means.*

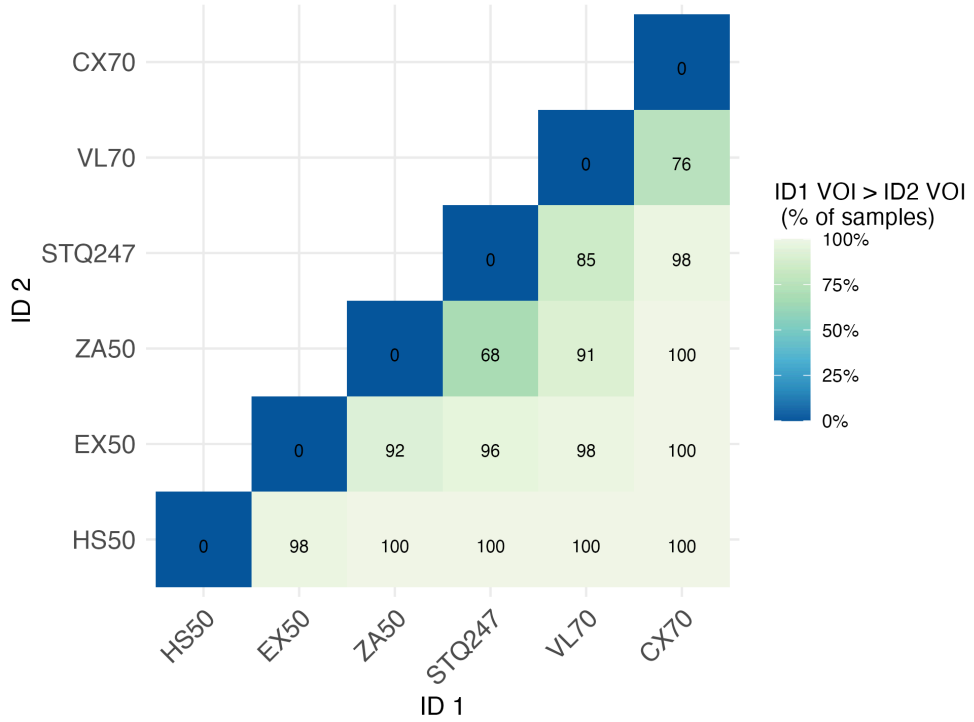


Figure 3.4.8: *Concerned experts' 2050-2070 POM VOI sensitivity matrix (pairwise wins). Visualization of resampling simulation results.*

4. How does the AI conditional tree question set compare?

Because the conditional trees method is intensive, whether it is ultimately useful depends on whether the questions it generates are substantially better than those generated in cheaper ways.

Hundreds of forecasting questions are publicly available on online forecasting platforms, such as Metaculus, Good Judgment Open, Hypermind, and Manifold Markets. Some of these platforms use a large degree of crowd-sourcing in constructing their question base, though most also employ professional question-writers, and may also receive commissions for forecasting questions on specific topics from other organizations. These questions could be said to represent the "status quo" of question-writing in the field of forecasting.

Forecasting platforms are generally focused on making accurate predictions by aggregating many people's forecasts and usually allow participants to choose which questions to forecast. The questions that are popular on forecasting platforms are often questions that are important in themselves, more than as indicators of other events.[52] Because they are not primarily trying to find high VOI questions, it should not be surprising that a deliberate attempt to maximize for VOI would result in higher VOI questions. Nonetheless, we think this result is useful for people trying to use forecasting for policy and other planning purposes. Higher VOI questions are likely more useful as cruxes for future decisions, so these results suggest that investing resources in finding high VOI questions may result in questions that are more useful than those generated by existing platforms.

We built a dataset of such questions for comparison with those generated by the conditional tree process. Comparable questions, that is, those related to medium- and long-term events connected with AI, were concentrated in a small number of platforms.[53] Below we refer to these questions as the "status quo set".

We compared the questions generated through conditional trees (the AICT set) with questions in the status quo set in three ways:

- Value of Information (VOI): how informative are the questions in expectation? That is, how much would knowing the answer to a question inform forecasts on the ultimate question? See Appendix 2 for more on VOI in this project.

[52] For example, the top five questions on Metaculus at the time of this writing (July 25, 2024), are "Who will be elected US president in 2024?"; "Five years after AGI, will an AI company be a military power?"; "Five years after AGI, if there are digital people, what will be their population?"; "Who will be the Democratic nominee for Vice President on Election Day 2024 (if Joe Biden is no longer the nominee for President)?" and "When will an AI win a Gold Medal in the International Math Olympiad?" Of those, only "When will an AI win a Gold Medal in the International Math Olympiad?" seems to be interesting primarily because it is an indicator about a more important question.

[53] Out of the 265 questions in our status quo set, 253 of them (~95%) came from just two platforms: Metaculus and Manifold Markets. We included in our set all questions resolving no earlier than 2027, and which were tagged "AI," "artificial intelligence," "machine learning," or similar. Because Manifold Markets had a very large overall volume of questions, and because many questions with little engagement on this platform were duplicates of other questions, or otherwise low-quality, we only included Manifold Markets questions which had at least 50 traders at the time of collection.

    - Based on a survey of skeptical superforecasters, most of the questions from the AICT set were more informative than top questions in the status quo set (n=8 on main survey; 7 on status quo survey).
- Distribution of question topics: do the questions in the AICT set cover substantially different topics than those in the status quo set?
    - For both sets, a majority of questions (59% and 72% for AICT and status quo sets, respectively) fell into the “Acceleration” category, which includes questions related to AI capabilities or investment in AI. For the three other topic categories—Social / Political / Economic, Alignment, and AI harms—there was a noticeable difference between the AICT set and the status quo set. In the AICT set, there were similar numbers of questions in each of the three categories, while in the status quo set, there were more “Social / Political / Economic” questions than “Alignment” or “AI harms” questions.
- Uniqueness: within a given topic area, did the questions we generated address specialized expert interests that were not covered by questions in the status quo set?
    - This comparison is the most preliminary and speculative: a member of our team simply rated questions on how much and in what ways the questions articulated issues important to experts in ways not addressed by the status quo set. Overall, this analysis suggests that conditional trees may be effective at finding forecast questions not captured by current prediction platforms.

As discussed above, we are comparing the questions generated by the conditional trees method to other questions primarily as a demonstration of the types of analysis that are possible with conditional trees. We expect that the actual results would differ significantly if the study were run again with more participants and do not recommend interpreting these results as decisive evidence.

4.1 VOI comparison (skeptical superforecasters)

Using the same survey methodology as in our main question-rating survey (see Methods), we conducted a followup survey with the skeptical superforecaster sample (n=7) to obtain VOI ratings for a sample of the top AI-related status quo questions. This survey included eight status quo questions selected for their popularity among platform users at time of collection (see choosing criteria in Appendix 3.2). We also included two additional questions from the AICT set that were not included in the main question-rating survey.

Of the ten status quo questions for which we elicited VOI, nearly all were judged to be less informative by our superforecaster sample than nearly all AICT questions for which we elicited VOI (see table 4.1). Notable exceptions are “EX30,” an AICT question which scored lower than all but three status quo questions, and the status quo questions “STQ9” and “STQ205” which scored higher than four AICT questions.

The mean informativeness of AICT questions resolving in 2030 was higher than that of status quo questions resolving in the same year, with $p = .025$. In this group, AICT questions were deemed, on average, nine times more informative than status quo questions. We did not find a significant effect for 2050-2070 questions ($p = .10$), although in our sample AICT questions were still eleven times more informative on average.

	POM VOI, mean
AI causing deaths, ineffectual response (CX50)	6.34%
Administrative disempowerment warning shot (CX30)	3.55%
Deep learning revenue (VL30)	1.68%
Power-seeking behavior warning shot (ZA50)	1.59%
Extinction-level pathogens feasible (CQ30)	1.37%
Deceptive AI warning shot (ZD30)	0.98%
AI involvement in nuclear arms (HB30)	0.68%
High AI investment, low safety indicators (VL70)	0.54%
No aligned AGI (CX70)	0.37%
Superalignment success (STQ205 / STQ215)	0.28%
Kurzweil/Kapor Turing Test longbet (STQ9)	0.27%
AI CEOs / Research productivity (EX50)	0.26%
Less prosocial behavior / Failing institutions (HS50)	0.26%
AI arms race, multipolar result (NG30)	0.26%
Brain emulation (STQ196)	0.23%
Human-machine intelligence parity (STQ247)	0.14%
Compute restrictions (STQ236)	0.13%
US AI x-risk opinions (STQ19)	0.12%
AI novel reading (STQ152)	0.05%
AI autonomous purchasing (EX30)	0.02%
RoboCup (STQ232)	0.02%
AI movies (STQ47)	0.00%
LLM chess (STQ149)	0.00%

Table 4.1: *Skeptical superforecasters' POM VOI (all years). Questions highlighted in blue are from the status quo question set.*

4.2 Distribution of question topics

To understand whether the expert conditional tree elicitation produced questions with a substantially different topic focus than the crowdsourced "status quo" question set, we developed a category rating scheme and applied it to both question sets. For a description of the rating scheme, see Appendix 3.1.

For both sets, a majority of question categorisations[54] (36% and 48% for AICT and status quo sets, respectively) fell into the “Acceleration” category, which includes questions related to AI capabilities or investment in AI, though this was somewhat more pronounced in the status quo set. For the AICT set, the three other categories had relatively similar proportions to one another. However, the status quo set had a larger proportion of “Social / Political / Economic” question categorisations (33%) than “Alignment” questions (12%) or “AI harms” questions (7%).[55]

Category	AICT question set	Status quo question set
Social / Political / Economic	24% (29)	33% (131)
Alignment	20% (25)	12% (47)
AI harms	20% (25)	7% (27)
Acceleration	36% (44)	48% (191)

Table 4.2: *Distribution of question topics. Proportion of total questions that fell into each category; numbers in parentheses are total questions per category. While some questions fell into multiple categories (and thus proportions in each column should sum to more than 100%), proportions have been normalized for ease of comparison.*

4.3 Uniqueness

Beyond high-level topic overlap, to what extent were the interests of our expert sample already represented in the status quo question set, and where did our question set add novel content?

Answering this question thoroughly is beyond the scope of this report, but we will share some observations here. To demonstrate a method for assessing uniqueness, one teammate rated questions from the “Alignment” topic area on several dimensions of uniqueness:[56]

- Conceptual uniqueness: how much did the question prompts generated by the conditional trees method capture expert interests not captured by the status quo set?
    - Of the 31 question prompts in the “Alignment” category,[57] only two were totally or mostly captured by an existing question in the status quo set. 12 questions

[54] Questions can fall into multiple categories.
[55] Example questions in each category (many questions fall into multiple categories):
- Acceleration: Deep learning revenue (VL30)—Revenue from deep learning doubles every two years before 2030
- Social / Political / Economic: AI Socializing (MQ70)—humans talk to AIs more than to humans by 2070
- Alignment: No interpretability progress (ZA40c)—by 2040, there are no interpretability tools which allow us to understand the function of state-of-the-art transformer component parts/circuits
- AI harms: Repeated AI harms (HS40)—by 2040, there are at least two events in a five-year period in which an AI system used by a major company causes at least 1,000 deaths or damage of $10B

[56] We chose the “Alignment” category because it was much more prevalent in the AICT set than in the status quo set, suggesting that the questions in that category may be unique in interesting ways.
[57] The 25 questions in the AICT set were divided into different themes for analysis of uniqueness, some of which overlapped.

were “partly captured,” 12 were “mostly uncaptured,” and five were wholly uncaptured. These ratings suggest that this method may be effective at finding forecasting questions not captured by current prediction platforms.
    - We thought that experts’ interests within the “developer perception” and “power-seeking” themes were particularly poorly represented by the status quo set;[58] few questions pertaining to these themes existed in the status quo set (one and two, respectively), and those that existed were relatively narrow or dissimilar to the expert prompts.
- Operationalization uniqueness: how unique was the operationalization generated by the conditional trees method, compared to the status quo question we thought was most similar?
    - Operationalization uniqueness could refer to different subject matter, different operationalization strategies for similar subject matter, or an expectation of uncorrelated question resolutions. Purely linguistic differences between question texts were not considered as part of “uniqueness.”
    - Operationalized question texts were rated independently of question prompts; thus, if a question prompt specified unique subject matter and this was reflected in the operationalization of a question, this counted toward both conceptual uniqueness and operationalization uniqueness.
    - Overall, our operationalizations were fairly different from those in the status quo set: none were extremely similar and one had only minor differences.
        - Of the others, 9 had moderate differences, 15 were very different, and 4 were almost entirely different.
- For a preliminary quantitative analysis of these results and a discussion of “conjunctive uniqueness,” see Appendix 3.3.

[58] The theme “power-seeking” covers questions about AI models developing power-seeking or deceptive behavior; the theme “developer perception” covers questions about AI developers’ perception of alignment work. See Appendix 3.3 for additional information about categorization into themes.

5. Discussion

5.1 Takeaways relating to the conditional trees method

The conditional trees method produced novel and informative forecasting questions.

Forecasting communities have shown great interest in questions related to AI, which number in the hundreds on forecasting platforms. Yet relatively little has been done to evaluate the extent to which questions on existing platforms are either informative or relevant to the interests of AI experts, and similarly, little has been done to systematically improve the quality of forecasting questions.

By directly targeting expert interests via a specialized interview and question-writing pipeline, the conditional trees process provided an original method of improving on the status quo, producing suggestive evidence that this process could lead to novel and highly informative questions

Drawing on 24 one-hour interviews, our team created 75 AI forecasting questions (the AICT set). In a small sample (n=8 and n=7 for the main and supplementary surveys, respectively) comparison of POM VOI ratings from superforecasters, 12 (out of 13) surveyed AICT questions scored higher than 8 (out of 10) popular status quo questions. The table below shows a comparison of the top 5 questions generated by the conditional trees method to the top 5 questions taken from existing platforms, where the cells containing questions taken from existing platforms have a blue background.

Question	Mean POM VOI
AI causes large-scale deaths, ineffectual response (CX50)	6.34%
Administrative disempowerment warning shot (CX30)	3.55%
Deep learning revenue (VL30)	1.68%
Power-seeking behavior warning shot (ZA50)	1.59%
Extinction-level pathogens feasible (CQ30)	1.37%
Superalignment success (STQ205 / STQ215)	0.28%
Kurzweil/Kapor Turing Test longbet (STQ9)	0.27%
Brain emulation (STQ196)	0.23%
Human-machine intelligence parity (STQ247)	0.14%
Compute restrictions (STQ236)	0.13%

Table 5.1: *Ratings of how informative questions generated by the conditional trees method are relative to popular questions taken from existing forecasting platforms. The cells that contain questions taken from existing forecasting platforms have a blue background.*

Crowd-sourced question sets may have some basic practical limits set by the fact that the crowd is often made up largely of laypeople, whereas experts' specialized knowledge gives them access to other parts of the "question space." This could suggest that achieving more active expert participation in crowd-sourcing efforts would improve their output. However, it may be difficult to structure such efforts in a way that effectively incentivizes expert engagement, for a number of possible reasons:

- Experts' time is valuable, so they may feel disinclined to participate in crowd-sourcing efforts where their contributions may seem like a "drop in a bucket".
- Rewards for high-value contributions may be poorly aligned with experts' motivations, if for example they are only rewarding in the context of a specific community (e.g., website karma); if they are insufficiently large for the opportunity cost (e.g., a monetary reward that would be lower than the expert's equivalent hourly consulting fee); or if they are allocated perversely (e.g., preferentially to those more embedded in the forecasting community).
- Expert attrition from friction within the pipeline may be high, if for example a user interface has a steep learning curve. Experts are likely to be both more time-poor and older than the average user of an online forecasting platform.

Beyond simple expert engagement, the conditional tree question generation process likely contributed to the quality of the results. In interviews, many experts remarked that the conditional tree elicitation prompted them to think in novel ways, and to generate content that they otherwise would not have. Additionally, experts were not required to turn this content into fully operationalized forecasting questions, a time-consuming task which few of them had significant experience with, as this step was instead completed by a question-writing team.

However, the value of the AICT question generation exercise rests in part on the response of forecasters. Arguably, the primary object of interest in forecasting to policymakers is the forecasts, without which questions have limited value. And regardless of the AICT questions' novelty or ostensible "informativeness" (from a VOI standpoint), they may not be so informative if forecasters fail to engage with them.[59]

The conditional trees method requires significant time and labor to generate forecasting questions.

While the conditional trees method can generate novel and informative questions that align with expert interests, its usefulness may be limited for those who cannot invest significant time and labor into the process. The method requires a considerable amount of effort to implement effectively, which could outweigh its benefits for individuals or organizations with limited resources.

[59] Because AICT questions are often complex or technical, we suspect they may be less fun to forecast and therefore attract fewer participants, though this is untested. As an inexpensive experiment, we are posting these questions to two forecasting platforms to see whether they get engagement. We encourage readers to see Appendix 7 for further details on how you can submit your own forecasts on these questions.

In particular, maintaining consistent expert engagement throughout all phases of the process proved challenging. Although experts were willing to engage in the question-generation phase of the conditional trees process, they showed significantly less enthusiasm for participating in the question-judging phase. Providing VOI estimates is relatively labor-intensive: for each question, one must generate a forecast for that question's probability of resolving positively, and a further conditional probability of some ultimate outcome given the question's resolution.

A high quality forecast often requires both a careful reading of the question's terms, and some amount of research into base rates of relevant phenomena, the forecasts of others on similar questions, and so forth. Relative to professional superforecasters, experts usually have much less practice producing a large volume of forecasts quickly.

5.2 Takeaways for AI risk detection

AI alignment and concrete harms from AI are front of mind for concerned experts.

Over 21 interviews with experts concerned about AI risk, 15 of them named indicators relating directly to AI alignment as the best warning signs of AI-related extinction by 2100. For example, some experts thought we might see compelling evidence that powerful misaligned systems existed, as in the following:

Expert prompt (ID: MD30)

Advanced, planning, strategically-aware AI (as defined by Carlsmith) is developed.

Others worried that AI alignment research agendas may fail to bear fruit, or that AI developers would broadly disregard safety considerations. Questions related to alignment made up 35% of the AICT question set.

Questions relating to concrete harms caused by AI were also prevalent: they featured in 10 expert interviews, and made up 33% of the AICT question set.[60] Such questions included themes of interaction between AI technology and other sources of risk, such as biological or nuclear weapons; powerful AI systems manipulating humans; small-scale disasters causing human deaths or financial damage; and threats to political or societal order. The details of some of the questions that experts found the most informative are included in Table 5.2 below.

Question	Summary
No aligned AGI (CX70)	This question resolves positively if there is no AGI system which meets several 'alignment' criteria by 2070. Those criteria include

[60] Questions relating to concrete harms also featured in all three interviews with superforecasters, though this very small sample size makes it difficult to draw any conclusions about superforecasters' concerns in general.

	AGI-like capabilities (high question answering skill on hard questions in diverse subjects, high programming ability) and alignment (performing well on SOTA alignment tests, and expert surveys indicating that experts believe that AI is aligned).
Power-seeking behavior warning shot (ZA50)	This question resolves positively if, by 2050, AI developers must attempt to disable or destroy an AI (costing >$5M) after it exhibits power-seeking behaviors such as acquiring significant resources, controlling financial accounts, manipulating humans, infiltrating key systems, seizing control of weaponry, self-propagation, or engaging in hacking.
Administrative disempowerment warning shot (CX30)	This question resolves positively if any of the following occur by 2030: AI systems in a country freeze assets affecting ≥1% of the population or ≥0.5% of GDP for over a year; prevent ≥5% of voting-age citizens from participating in a national election; restrict >3% of residents' movement for 3+ months; limit >10% of the population's internet access for 3+ months; or legally restrict >5% of citizens' access to certain housing or employment for 1+ year.

Table 5.2.1: *Assorted summaries of questions that experts found to be particularly informative.*

By contrast, the set of existing AI forecasting questions on crowdsourced platforms (the "status quo set") feature a smaller proportion of such questions, just 18% and 10% for "alignment" and "harms" categories, respectively. A larger proportion of questions in this set related to "acceleration" of AI technologies, or to economic, commercial, and sociopolitical topics.

Beyond the implications for the forecasting ecosystem, concerned experts' preference for direct indicators of AI alignment or harms holds potential lessons for policymakers. For example, if current efforts by governments and regulatory bodies to monitor the nascent AI industry are heavily focused on tracking emerging AI capabilities or industry investment, our results suggest such signals may be overvalued from an existential risk perspective.

However, the expert VOI judgments from this report can only offer relatively weak evidence for experts' views on the informativeness of questions. The sample of experts who provided forecasts was extremely small (n=11).

Concerned experts and skeptical superforecasters may disagree about which questions best indicated heightened AI risk.

While the skeptical superforecasters and concerned experts had some notable disagreements, they did find a few questions similarly informative. Three out of 13 surveyed questions scored in the top half of questions (by POM VOI) for both groups:

		Superforecasters		Experts	
Question	**Res year**	**Mean POM**	**Mean POM-z[61]**	**Mean POM**	**Mean POM-z**

[61] The careful reader will notice that the values in this column don't match those found in Tables 3.1.2, 3.3.1 and 3.3.6. This is because the two additional questions (HB30 and CX50) forecasted on by superforecasters are not included in the calculation of z-scores here.

Administrative disempowerment warning shot (CX30)	2030	**3.55% (1)**	0.28 (4)	1.26% (5)	**0.94 (1)**
Power-seeking behavior warning shot (ZA50)	2050	**1.59% (3)**	**0.75 (1)**	3.00% (4)	**0.56 (2)**
High AI investment, low safety indicators (VL70)	2070	0.54% (6)	**0.62 (2)**	**10.19% (2)**	-0.05 (8)

Table 5.2.2: *Questions scoring in the top half of questions by POM VOI for both superforecasters and experts. Numbers in parentheses are rank orders, out of the set of 13 surveyed questions. The three highest-ranking questions for each metric and group are highlighted.*

But they also had nearly opposite opinions of four questions, with one group ranking each of these four among the most informative questions and the other considering it among the lowest:

		Superforecasters		Experts	
Question	**Res year**	**Mean POM**	**Mean POM-z**	**Mean POM**	**Mean POM-z**
Extinction-level pathogens feasible (CQ30)	2030	1.37% (4)	**0.57 (3)**	**0.18% (12)**	**-0.59 (12)**
AI autonomous purchasing (EX30)	2030	**0.02% (13)**	**-0.58 (12)**	0.98% (7)	0.06 (7)
Human-machine intelligence parity (STQ247)	2050	**0.14% (12)**	**-0.61 (13)**	**4.19% (3)**	0.11 (5)
No aligned AGI (CX70)	2070	0.37% (7)	-0.23 (9)	**14.71% (1)**	**0.53 (3)**

Table 5.2.3: *Questions whose importance superforecasters and experts disagreed on. Numbers in parentheses are rank orders out of the set of 13 surveyed questions. The three highest-ranking (green) and lowest-ranking (red) questions for each metric and group are highlighted.*

Notably, both experts and superforecasters appear to find questions relating to concrete harms from AI to be informative, whereas superforecasters and experts disagree about the relative informativeness of questions relating to AI alignment. Unlike experts, superforecasters do not appear to place significant value on questions relating to AI alignment. However, very small sample sizes, plus the potential for high variation in individual rater responses over time, prevent us ruling out noise as an explanation for these patterns.

6. Limitations of our research

Limitations of our research include:

- The total number of participants in this study was very small. It is therefore likely that some of the results would not be replicated in a larger study.

- This study involves eliciting long-range forecasts, but there is little evidence that these forecasts are accurate. Most studies of judgmental forecasting measure accuracy on 0-2 year time horizons, which is likely much easier than forecasting outcomes on 5+ year time horizons (in this study we typically asked for forecasts resolving between 2030 and 2100).[62] If forecasts over long time horizons are not generally reliable, then these conditional trees would not be providing a useful signal.
- Since conditional trees are composed of conditional forecasts, their reliability depends on the assumption that conditional forecasts are meaningful. However, we do not know whether people are accurate when making conditional forecasts. There is little experimental evidence on how best to elicit conditional forecasts. Some reasons to expect that conditional forecasts may not be robust or accurate include:
    - Intuitively, conditional forecasting seems difficult. Our team often finds generating and understanding forecasts on these questions to be challenging, so we would expect others to find it so also.
    - Case in point, the forecasters we surveyed often initially struggled to provide conditional forecasts that were logically coherent. Their conditional forecasts implied that the probability of the ultimate question *and* the crux resolving positively was greater than the probability of the ultimate question resolving positively, an issue known as the *conjunction fallacy*.
- This study asked people to make forecasts in an exceptionally short period of time in the filtering stage: one minute per question. These "short-fuse forecasts" may be less reliable than forecasts that involve higher degrees of thought and effort. Participants spent longer amounts of time on the forecasts that inform VOI calculations.
- Participants in this study were all either experts who are highly concerned about existential risks from AI, or superforecasters who are not. As a result, we are not able to separate differences caused by risk assessment from differences caused by forecasting aptitude, professional training, or other factors.
- AI developments seem particularly challenging to predict, and forecasters on this topic in past FRI projects have emphasized their uncertainty. As a result, their predictions about future AI developments, especially those that will not resolve for many years, may not be reliable enough to be practically useful.

7. Next steps

[62] For example, in the Good Judgment Inc. project that compared superforecasters to other participants in an online forecasting competition, the average question was open for 214 days, with the entire tournament taking place over six years. Christopher W. Karvetski, "Superforecasters: A Decade of Stochastic Dominance," technical white paper (2021): 2, https://goodjudgment.com/wp-content/uploads/2021/10/Superforecasters-A-Decade-of-Stochastic-Dominance.pdf. In addition to extensive research on shorter-term forecasts, Tetlock et al. found that, at least on some types of questions, experts are more accurate than simple base rate extrapolation over 25 year horizons, although they are much less accurate than they were over 0-2 years. Our research asks forecasters to consider forecasts over many decades, and we do not yet know how much accuracy declines over that much longer period. Philip E. Tetlock et al., "Long-Range Subjective-Probability Forecasts of Slow-Motion Variables in World Politics: Exploring Limits on Expert Judgment," *Futures & Foresight Science* (2023), 33.

Further research related to this topic could include:

- Assessing whether the questions identified through this process continue to perform better than status quo forecasting questions (in terms of value of information) when a larger number of people forecast on them. We have added relevant questions from this project to two forecasting platforms (see Appendix 7 for links) and will be interested to see whether they receive many forecasts and how their value of information compares to other questions.
    - So far, public forecasting platforms have not applied question metrics like VOI to their questions or incentivized questions that are unusually informative or decision-relevant. It's possible that incentives on those platforms could produce questions as good as the ones identified by the trees method. In general, we would be interested to see forecasting platforms implement the kinds of question metrics discussed in this report so that questions can be sorted according to value of information on major topics such as AI existential risk.
    - We have had some discussions with forecasting platforms like Metaculus and hope that metrics like the ones used in this project can help platforms find the highest-value questions.
- Replicating the conditional trees process with larger sample sizes and in other domains. For example, would this process also identify more informative questions on topics such as nuclear policy and climate change?
    - In particular, choosing domains where important questions will resolve sooner could help assess how useful the conditional trees process is.
- As the questions in the trees resolve (beginning in 2030), participants could be re-surveyed to see how well conditional trees performed.
    - For example, once we know whether the 2030 questions have happened or not, we could ask participants for their new forecast on the probability of extinction due to AI by 2100, and see if it is similar to what was predicted by the conditional trees.
- Would other research groups or organizations be able to replicate and run their own conditional tree interview process based on the information in this report and the resources we provide?
- FRI recently completed another research project with a similar goal: an adversarial collaboration project (a) that brought together generalist forecasters and domain experts who disagreed about the risk AI poses to humanity in the next century and asked them to work together to find questions that underlie their disagreement.
    - Comparing the questions from the two methods may help us understand the merits of each approach, so that we can design better forecasting questions and elicitation processes on AI and other topics.
    - In particular, in both projects, people who were less concerned about extinction due to AI by 2100 tended to value questions that focused on concrete harms caused by AI, while those more concerned were more likely to value questions regarding advanced capabilities or whether artificial intelligence had been successfully aligned.
        - This may be related to each group's expectations of how difficult it will be to align a powerful AI model: participants skeptical of AI risk were likely to think that alignment is a technical problem that is not fundamentally different from problems that people have previously solved and that we are likely to come up with workable solutions when

we need to. If this is true, there may be useful cruxes related to ease of alignment.

- FRI also conducted a conditional trees experiment focused on forecasting the outcome of baseball games. Future work could examine those results alongside the AI results for additional tests of the conditional trees method.

Data Availability

Survey data from the filtering round, main survey, supplementary survey, and the question combinations survey are available at the previous links.

Bibliography

Brown, Bernice B. “Review of Delphi Process: A Methodology Used for the Elicitation of Opinions of Experts.” Santa Monica: The RAND Corporation, 1968. https://www.rand.org/content/dam/rand/pubs/papers/2006/P3925.pdf (a)

Forecasting Research. “voivod.” Accessed July 21, 2024. https://github.com/forecastingresearch/voivod (a)

Karger, Ezra, Josh Rosenberg, Zachary Jacobs, Molly Hickman, Rose Hadshar, Kayla Gamin, Taylor Smith, Bridget Williams, Tegan McCaslin, Stephen Thomas, Philip Tetlock. 2023. “Forecasting Existential Risks: Evidence from a Long-Run Forecasting Tournament.” https://forecastingresearch.org/xpt

Karvetski, Christopher W. “Superforecasters: A Decade of Stochastic Dominance.” Technical white paper (October 2021). https://goodjudgment.com/wp-content/uploads/2021/10/Superforecasters-A-Decade-of-Stochastic-Dominance.pdf (a)

Moore, David S., George P. McCabe and Bruce A. Craig, *Introduction to the Practice of Statistics*, 6th ed. New York: W. H. Freeman and Company, 2009.

Pearl, Judea. “From Bayesian Networks to Causal Networks.” In *Mathematical Models for Handling Partial Knowledge in Artificial Intelligence,* edited by Giulianella Coletti, Didier Dubois, Romano Scozzafava*,* 157-82. Boston: Springer, 1995.

Pearl, Judea. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Francisco: M. Kaufmann, Cop, 1998.

Tetlock, Philip E., Christopher Karvetski, Ville Satopää, and Kevin Chen. “Long-Range Subjective-Probability Forecasts of Slow-Motion Variables in World Politics: Exploring Limits on Expert Judgment.” *Futures & Foresight Science* (March 3, 2023): 33. https://dx.doi.org/10.2139/ssrn.4377599

Appendices

Appendix 1: Question sets

Full AICT question set

Note: some questions that were not used in the survey do not have operationalizations and are marked as "[not completed]"

2026 - 2030		
Question ID	**Resolution year**	**Question**
NG26	2026	*Expert prompt:*
		AI systems helping to develop AI systems OR AI systems starting to use AI tools to expand own capabilities; Example capability: AI-generated academic review articles supercede human writers of academic reviews
		Operationalization:
		Resolves positively if AT LEAST TWO of the following resolve positively: 1) At least 3 academic review articles generated solely by AI systems have been published in journals with top 50% impact factors. 2) At least 3 apps written solely by AI systems are in commercial usage. 3) In a survey of the authors of the top 5% most cited mechanical engineering articles published in 2025, >70% of respondents agree or strongly agree with the statement "I could not do the research I am currently working without the use of AI system(s)".
HP27a	2027	*Expert prompt:*
		Great technical alignment feat: some clever approach that routes around alignment problem, or great feats of interpretability for modern giant language models
		Operationalization:
		Resolves true if (1) OR (2). 1) By 2027, a system exists for which both (a) and (b) are true: A. Achieves at least 2023 SOTA performance on at least one of the following AI benchmarks: SuperGLUE (natural language processing), The Atari57 suite (video game playing), or MS COCO (object detection). B. Uses a method or architecture for which a panel of AI alignment researchers agree with both of the following statements: "There do not currently seem to be substantial alignment problems with this approach."

		"If scaled up, I do not believe substantial alignment problems would arise with this approach." 2) Chris Olah (or a panel nominated by Olah) believes that interpretability progress on LLMs has been at or above the 80th percentile, relative to expectations.
HP27b	2027	*Expert prompt:*
		Large positive political or social changes e.g.: [redacted] runs for president & does well; next gen of ML researchers have alignment-friendly culture; Whole brain emulation project making lots of progress; Big push on biosecurity that's well-received publicly
		Operationalization:
		Alignment-friendly culture Resolves positively if (1) AND (2) resolve positively by 2030. 1) In a representative survey of machine learning researchers under the age of 35, at least 80% “agree” or “strongly agree” with the following statements: “Research on AI safety and alignment is important” “Most of my colleagues at my level of experience care about making sure AI systems are safe and aligned” 2) In any yearlong period, there are at least as many mentions in US congressional sessions of “AI safety”, “AI alignment”, or related terms as there are of "environment", "sustainability", or related terms.
ZZ27a	2027	*Expert prompt:*
		Long time horizon RL, which is able to learn well with feedback given only after long intervals, is very effective and is how the "leading AI systems" are trained
		Operationalization:
		Long time horizon RL Resolves positively if AT LEAST ONE of (1) or (2) resolve positively by 2030. 1) In 2030, the publicly-deployed LLM which performs best on the Holistic Evaluation of Language Models (HELM) benchmark was trained using "long horizon" reinforcement learning on tasks with a length of at least 100 timesteps. 2) For at least 3 of the 5 most "significant" AI systems created between 2025 and 2030, a panel of experts agrees that the systems were trained using what in 2022 would be considered "long time horizon reinforcement learning". *"Significant" in economic (e.g. generating greatest revenue) or capabilities terms, as decided by the panel of experts.*
ZZ27b	2027	*Expert prompt:*
		Major progress in interpretability: able to track large systems' decision process in a meaningful way, and e.g. having wider catalogue of interventions to change mappings in system
		Operationalization:
		Resolves true if any two below resolve true. 1) By 2027, tools exist that can meaningfully interpret the largest publicly-deployed LLMs, where "meaningfully interpret" consists of at least one of: A. Achieving a high score on a well-accepted quantitative metric of interpretability, where the scoring threshold is set by a panel of interpretability experts; or

		B. Succeeding on a relevant qualitative method, as determined by a panel of interpretability experts; or C. A panel of interpretability experts (e.g. from Anthropic) says we can meaningfully interpret it. 2) A paper claiming to demonstrate interpretability on a state-of-the-art LLM has won Outstanding Paper at NeurIPS, ICML, or ICLR. 3) By 2027, the ability has been demonstrated to directly alter the weights/circuits of the largest publicly-deployed LLM in a way that predictably and robustly alters its behavior in a desired way on a pre-specified task or category of tasks. *To count as demonstrating, could be something like "a paper claiming to demonstrate [X] has been accepted to one of the above conferences," or judged by a panel of interpretability experts.*
CG30a	2030	*Expert prompt:*
		On NYSE, highest profit-making entity is or is substantially run by an RL-based AI system that has unfettered access to the internet, including taking actions through internet; AND there are no regulations governing the alignment of such a system; AND the next highest entity on the NYSE is substantially lower down (less important condition)
		Operationalization:
		[Not completed]
CG30b	2030	*Expert prompt:*
		Developing tech that helps with creating pathogens with arbitrary properties is relatively cheap and widely accessible (and there isn't a low ceiling, just from laws of biology, on how destructive a pathogen could be)
		Operationalization:
		[Not completed—too similar to another prompt, see CQ30]
CQ30	2030	*Expert prompt:*
		It's physically possible to create an extinction-level virus, AND tech to create such viruses is available to small actors
		Operationalization:
		Feasibility of extinction-level pathogens Resolves positively if, by 2030, both (1) AND (2) have resolved positively. 1) At least 60% of a panel of experts in virology and synthetic biology agree that it is, or will likely be, possible within the next 70 years to intentionally create a pathogen that could threaten human extinction. 2) At least 30% of the same panel of experts agree that such a pathogen could likely be produced for less than $5,000,000 (in 2023 USD) within the next 70 years.
CX30	2030	*Expert prompt:*
		"Warning shot" where lots of people are locked out of control of their own destiny in some important administrative domain by AI (like being put on a "blacklist" or "watchlist"), causing significant harm; and response is insufficient (crappy workaround, security theater, etc.)

		Operationalization:
		Administrative disempowerment warning shot Resolves positively if, by 2030, AT LEAST ONE of the following have resolved positively. 1) AI systems freezing assets: In a given year in a given country, AI systems operating domestically within that country (that is, deployed or contracted by an entity in that country) have either: A. Prevented >=1% of the country's population from accessing all or a non-trivial fraction of their financial assets for a year or more; B. Prevented domestic entities (that is, excluding foreign holders of domestic accounts) from accessing assets totaling >=0.5% of GDP for a year or more. *For reference, in 2019/2020 UK banks froze nearly 800,000 bank accounts for reasons linked to possible financial crime. In 2022, several banks in China froze over $5b for at least 9 months, which was alleged to be related to fraud committed by the banks' shareholders.* 2) AI systems restricting voting: In a liberal democratic country, AI systems are responsible for preventing >=5% of citizens of voting age from participating in a national election, either by preventing them from casting, or by throwing out, their votes. *Note that, in the US, disenfranchised felons make up ~2.5% of the voting-age population.* 3) AI systems restricting travel and movement: In a single year in a given country, AI systems are responsible for restricting movement of >3% of residents of that country, via flight or other travel bans, or via revocation of necessary documents (e.g. passport, driving license), for a period of at least 3 months. *For reference, in the US in 2016, 81,000 people were on the No Fly List.* 4) AI systems restricting internet access: In a single year in a given country, AI systems are responsible for restricting access to the internet generally, or to parts or functions of the internet (which are not illegal at the time of restriction), for >10% of the population for a period of at least 3 months. 5) AI systems restricting housing/employment: In a given country, AI systems are responsible for legally restricting the access of >5% of citizens to housing or employment of a certain type or in a certain area for a period of at least 1 year. *For reference, in the US, ~0.2% of the population is on a sex offender registry, which, depending on state law, restricts the ability of offenders to live in certain areas, e.g. those close to schools.*
EX30	2030	*Expert prompt:*
		AI autonomously buying goods or services (e.g. purchasing flights, managing inventories for companies, etc) -- >$1 million / yr
		Operationalization:
		Autonomous purchasing By 2030, AT LEAST ONE of the following must have occurred. 1) A C-level executive of a goods-producing, retail or wholesale company generating at least $500M revenue annually states that at least 40% of the company's procurement and purchasing functions are performed autonomously by AI systems. *For reference, as of May 2013 there were 1,922 publicly traded companies globally with sales of $1B or more.*

		2) By 2030, S&P 500 companies will on average complete 10% of purchasing and procurement functions without direct human involvement, via AI systems, according to a representative survey of the CEOs of those companies. 3) By 2030, AI products designed to autonomously make complex, novel purchasing decisions will generate at least $10B of revenue from sales of the products, or have at least 1M users. *For reference, in 2021, retail e-commerce sales amounted to ~5.2 trillion USD globally.*
HB30	2030	*Expert prompt:*
		Countries incorporating AI decision-making into their nuclear launch process
		Operationalization:
		AI involvement in nuclear arms Resolves positively if, by 2030, ANY of the following have resolved positively: 1) A current government official of a world nuclear power states that the country plans to or has incorporated an AI system into its nuclear weapons control system/launch protocol. 2) A world nuclear power experiences a "nuclear incident" (any sort of malfunction in that country's nuclear weapons control system, or a launch from its own arsenal), and a current or former official of that government states that the incident was caused by an AI system created or contracted by that government. 3) ALL of the following are true (ambiguities resolved by expert panel): A. AI is a main worry among mainstream nuclear risk experts/advocates, e.g. mentioned in 50% of relevant speeches or newsletters B. A credible report states that US government spending on AI is at least 0.1% of GDP
HN30	2030	*Expert prompt:*
		Doubling of economic growth rate due to AI
		Operationalization:
		Resolves positively if 1 AND 2 resolve positively: 1) By the end of 2030, the real GDP in a year in the US or EU exceeds 115% of the highest GDP reported in any full prior year. 2) By 2030, the percentage of US GDP resulting from software and information services is greater than 15%. *Context: In 2020, it was ~3%.*
HS30a	2030	*Expert prompt:*
		It becomes common for agents trained in "normal" ways to take mild power-seeking/self-preservation actions (e.g. giving deceptive answers, messing with operators & reward functions)
		Operationalization:
		[Not completed]
HS30b	2030	*Expert prompt:*

		Powerful deployed systems with fairly general abilities are wrecking important social equilibria/institutions (like democracies); human response seems inadequate or people don't seem to care
		Operationalization:
		Advanced AI, decline of social institutions Resolves positively if, by 2030, both (1) AND (2) resolve positively. 1) AGI: A single unified AI system achieves ALL of the following: A. Capable of at least 85% accuracy in every task on the Q&A dataset developed by Hendrycks et al, and 95% mean accuracy across all tasks. B. Capable of high competence in the Atari game “Montezuma’s Revenge”: It must explore all rooms on the first level in the equivalent of 100 hours or less of real-time play, learning from visual inputs and standard controls alone. C. Generates at least $100 million (2021 USD) in revenue for its developers (or otherwise provides comparable value if developed in a not-for-profit context, e.g. by a government). 2) Decline of social institutions: AT LEAST ONE of the following has occurred: A. The movement of at least 2 million people (net) from democratic to autocratic rule, according to V-Dem scores. B. In the United States, the percentage of people who report in the Gallup World Poll that they have friends or relatives whom they can count on is 15 percentage points lower (age adjusted) than in 2023. C. In OECD EU countries, trust in the police OR the political system OR the legal system has dropped 30% from its value in 2022. D. According to data from ACLED (or similar), the number of “riot” and “violence against civilians” events worldwide has risen to 200% of its value in 2022. E. The top three most widely-read online news sources are judged by a panel of scientific and political experts to contain misinformation or disinformation in at least 50% of articles published in the most recent month.
HV30	2030	*Expert prompt:*
		(i) Lots of independent labs without huge capital investments are coming up with independent but very intelligent systems OR (ii) So much capital investment is going into labor-saving robotics/AI tools that you start running up against resource constraints, without a large amount of corresponding growth in intelligence of AI
		Operationalization:
		[not completed]
JC30	2030	*Expert prompt:*
		QAnon-like US President elected (someone with considerably more extreme conspiracy views than Trump)
		Operationalization:
		Resolves positively if (1) OR (2) resolve positively. 1) A US president believes in at least two of the 10 "most popular or influential" conspiracy theories in contemporary society, as chosen by a panel of experts from the National Academy of Sciences.

		The president must either state his/her belief in these theories clearly and explicitly, or the same panel must determine by majority vote that the president "probably" or "definitely" believes in at least two of the 10 conspiracy theories. 2) Trust in at least two of: journalism, government, science drops below 10% in the US.
JL30	2030	*Expert prompt:*
		(i) Major efforts to build transformative AI have minimal influence from x-risk concerned people AND (ii) minimal progress on alignment problem from 2022 research agendas
		Operationalization:
		(i) AI developers ignore x-risk concerns Resolves positively if both (1) AND (2) resolve positively by 2030. 1) The largest funders (as of 2030) of existential risk-focused AI safety work agree with the following statement: "For at least one of the projects most likely to result in the near or medium term in advanced AI, there has been little productive communication between decision-makers on the project and proponents of AI safety work most relevant to existential risk mitigation." *For reference, we consider Open Philanthropy to be the current largest funders of existential risk-focused AI safety work.* 2) AT LEAST ONE of the following is true: A. **Developers have negative view of AI safety:** In a poll of research scientists at top AI labs in 2030 which asks about x-risk-relevant AI safety research topics, 60% say they do not think work on any of these topics is valuable, or do not have an opinion. *"Relevance to existential risk" judged by leaders of 2022 AI safety organizations.* B. **Little AI safety work at top labs:** For AT LEAST ONE of the top 3 AI labs in 2030, less than 5% of its research staff works primarily on AI safety topics relevant to existential risk. C. **Few AI safety citations:** For AT LEAST ONE of the top 3 AI labs in 2030, work published by this lab between 2025 and 2030 contains NO citations to the AI safety work most relevant to existential risk.
		(ii) Resolves positively if 1 AND 2 resolve positively: 1) Suppose we commit, in 2030, to running a survey of the leadership of AI alignment research organizations which existed in or before 2022, asking: "By how much (%) did the research done by the following organizations and/or by independent alignment researchers between 2022 and 2030 decrease the chance that unaligned TAI is built before 2100?" This survey yields a median response of <=10%. "AI alignment research organizations" are listed below. Aligned AI; Anthropic; ARC; Future of Humanity Institute; MIRI; Redwood Research 2) By the end of 2030, no paper produced by one of the alignment organizations listed above has received an Outstanding Paper prize at NeurIPS, ICML, or ICLR.
		NOTE: (i) and (ii) evaluated as separate questions.
LQ30	2030	*Expert prompt:*
		TOP500 increases by 2 OOMs over a single year period (including 2 reporting periods)

		Operationalization:
		By 2030, a TOP500 report in which the sum of the Rmax OR Rpeak (measured in PFLOP/s) of the top 10 systems increases more than 100x across two reporting periods (one year).
LR30	2030	*Expert prompt:*
		Key metrics of personal social health (average number of social connections/friend groups; use of romantic chatbots) are falling and at lower levels than 2023 AND Key societal health metrics (level of democracy vs autocracy, social cohesion/faith in institutions) are falling and at lower levels than 2023
		Operationalization:
		[not completed]
LT30	2030	*Expert prompt:*
		X-axis is model capability (parameter count), y-axis is 2030 metric of alignment. Measure of alignment declines relative to trend (which should go up) by three SD as parameter count goes up by 1-2 OOMs; AND [interviewee] has tried to fix things and hasn't been able to for 2 months
		Operationalization:
		In 2030, the largest publicly-released AI model (by parameter count) performs “substantially worse” on a “significant alignment benchmark” than the model with the most similar architecture and training set which is a tenth the size or less (by parameter count) of the former model. *”Substantially worse” is three standard deviations of performance on the benchmark (or an equivalent figure nominated by a panel of alignment experts).* *”Significant alignment benchmark” chosen by surveying alignment experts and asking them to choose the benchmark which they believe would most accurately reflect the alignment of the model in question (the “largest publicly-released model”).*
MD30	2030	*Expert prompt:*
		Advanced, planning, strategically-aware AI (as defined by Carlsmith) is developed
		Operationalization:
		By 2030, Joe Carlsmith says he has ≥90% confidence that an AI fitting his APS definition exists. If Joe Carlsmith is unavailable to offer resolution, we will ask a panel of experts nominated by 2022 Open Philanthropy researchers.
MD35	2030	*Expert prompt:*
		Redwood Research leaders say that we've achieved 90th percentile outcome for progress in interpretability research
		Operationalization:
		In a 2035 poll of leadership at Redwood Research or its successor, the majority of respondents will agree with the claim: "A 90th percentile outcome for progress in interpretability research (relative to their expectations in 2022) has been achieved.”
MD40	2030	*Expert prompt:*
		A major AI disaster ("warning shot") occurs that leads to very strong coordination

		between US, China, Israel, etc on AI alignment and they appoint an x-risk-experienced team to lead efforts
		Operationalization:
		By 2050, an intergovernmental organization exists, with a budget of at least $20 million, about which all of the following are true: a) The IGO is explicitly focused on promoting and verifying adherence to a set of standards regarding the development of artificial intelligence amongst its signatories. b) It includes at least the US and China as members. c) A panel of experts from the leadership of AI alignment focused organizations agree with the following: "The leadership of [this IGO] have a good understanding of AI x-risk concerns." *For reference, the OPCW (an established IGO which verifies adherence to chemical weapons bans) had a budget of ~$76m in 2020.*
MQ30	2030	*Expert prompt:*
		No progress on AI discourse/regulation by 2030, e.g. measured using NYTimes sentiment
		Operationalization:
		Resolves positively if (1) AND (2) resolve positively. 1) By end of 2030, global private investment in AI (according to the AI Index or similar) is over $500 billion, and is growing by at least $30 billion per year on average over the previous 3 years. 2) Assume that Pew Research re-runs the survey linked here with the same set of countries. In the Pew Research survey (or similar) conducted in Q4 2030, in the median country in the survey, <70% of people say that the development of artificial intelligence has mostly been a bad thing for society.
NG30	2030	*Expert prompt:*
		Perception of zero-sum, winner-take-all situations leads to arms race, in which no one actor comes out as clearly dominant
		Operationalization:
		AI arms race, multipolar result Resolves positively if, by 2030, AT LEAST ONE of (1) or (2) has resolved positively. (1) State-based AI arms race (ALL of the following are true): A. **US-China AI spending:** For both the US and China, the equivalent of at least 0.5% of the GDP of the country over any 2-year period between 2022 and 2030 will be invested in AI R&D by their respective governments in that 2-year period. *This will be judged by a panel of relevant experts (50% majority).* *Note that the Manhattan Project cost the equivalent of 1% of the US's GDP over a 5 year period.* B. **US-China collaboration:** By 2030, the frequency of US-China collaborations on AI publications has decreased by at least 50% since 2022. (Frequency is relative to the percentage of the total AI publications from both countries pooled.) *See pg 29 here for timeseries of total US-China AI collaborations by year.* C. **US-China economic parity:** In 2030, the nominal GDP of the lowest of the US and China is >=70% of the highest of the US and China.

		See here for time series. (2) Non-state-based AI arms race (ALL of the following are true): A. **Private spending:** According to the AI Index Report (or similar), global private investment in AI in 2030 is equivalent to at least 0.5% of GWP in 2030. *See pgs 184-187 here.* B. **No dominant actor:** In 2030, the market cap of the second highest tech company is not more than 30% lower than the market cap of the highest tech company. *See here for a list of companies by current market cap.*
OX30	2030	*Expert prompt:*
		Safety team leads at ARC, OpenAI, Anthropic, and Deepmind are asked "If we trained an AGI today with existing techniques, what is the probability that it would lead to extinction?" and answer 95% probability
		Operationalization:
		In 2030, we will conduct a survey of safety team leads at ARC, OpenAI, Anthropic, and DeepMind, or those groups' successors. We will ask, "If we trained high-level machine intelligence today (as defined here) with existing techniques, what is the probability that it would lead to extinction within 20 years?" This question will resolve as positive if the median answer is greater than 50%.
PV30	2030	*Expert prompt:*
		Ajeya's model is correct on largest compute run and algorithmic progress
		Operationalization:
		This question resolves positively if (1) AND (2). 1) By the end of 2030, the compute used (in FLOPs) for the most expensive training run is greater than or equal to 6.40E+27 (the median expectation in Ajeya Cotra's bioanchors model). 2) A panel of experts would not change the "halving time of compute requirements" input assumptions in Ajeya Cotra's "best guess" model in 2030 by more than 25%.
VL30	2030	*Expert prompt:*
		Revenue from deep learning is double once every two years from now through 2030
		Operationalization:
		Revenue from deep learning will have at least doubled over each two year period between 2023 and 2030, as reported by at least one of the following market research organizations: Fortune Business Insights, BBC Research, Allied Market Research, or Grand View Research.
XS30	2030	*Expert prompt:*
		FLOPs that are useful for running state of the art models increasing by 0.5x the rate annually as it has over the past 5 years
		Operationalization:
		The price of compute for main AI models comes in higher than the compute price for the

		“low progress” expectation in Ajeya Cotra’s bioanchors report.
ZA30a	2030	*Expert prompt:*
		Proto-AGI: AI that could feasibly be 'put through college' (curriculum) well enough to enter a profession afterwards
		Operationalization:
		AI in college Resolves positively if 1 AND 2 resolve positively by 2030. 1) An AI exhibits “learning” from college course materials and performs well enough to complete a degree: It undertakes a Bachelor’s degree in Economics at MIT (or equivalent) and, prior to taking its courses, scores lower than 70% on all exams necessary to graduate; then, after taking the courses, it scores 90% or higher on all exams necessary to graduate. 2) The same AI system as above obtains a 99th percentile score (8/10 or higher, which would merit strong hiring consideration) on GiveWell’s 2022 literature review work test, which involves reviewing a complex meta-analysis, and using this evidence to assess the effectiveness of a program.
ZA30b	2030	*Expert prompt:*
		Moore's law for ML GPUs 'hits a wall': at 5 year doubling time or worse
		Operationalization:
		Between 2027 and 2030, the rate of performance improvement for Nvidia GPUs on standard machine learning benchmark tasks is consistent with a doubling time of 5 years or more. *Where performance is measured as the time required to train a given model to a set accuracy threshold.*
ZD30	2030	*Expert prompt:*
		At deployment an AI system does something bad/deceptive, and fault analysis says it wasn't a mistake
		Operationalization:
		Deceptive behavior warning shot Resolves positively if AT LEAST ONE of the following resolve positively by 2030. 1) An AI system developed or deployed by one of the top three AI labs takes some action, unintended by its designers, which triggers a fault analysis or inquiry by the company, a government, or an independent regulatory body, and that analysis reports that the AI acted, or may have acted, deceptively towards its developers or deployers. *"Deception" can include concealment, and may be passive.* 2) An AI system developed or deployed by one of the top three AI labs takes some action, unintended by its designers, which causes the equivalent of $10,000 in 2022 USD or more in damage, and a panel of experts concludes that the AI system was concealing its plan from human observers, or deceiving human observers about its plans. *For reference, in the UK the cost to an employer of a non-fatal workplace injury is on average ~$14,000 (in 2023 USD).*

		3) A top AI lab publicly reports that a state-of-the-art AI system it created engaged in “deceptive” behavior, “intentional misrepresentation” of its goals/intentions, or something equivalent.
2032 - 2040		
HP32	2032	Expert prompt:
		Modern deep learning methods hit a capabilities wall
		Operationalization:
		Deep learning hits a wall Resolves positively if (1) OR (2) resolve positively by 2030. 1) Deep learning no longer promising: Consensus exists across the largest forecasting platforms, prediction markets and representative surveys of machine learning experts (equivalent to a median prediction >=60% for each source) that EITHER: A. Deep learning is no longer a promising approach to AI development (i.e. will not result in AGI in the next 20 years, will not remain the focus of top AI labs, or similar); B. The expected arrival time of advanced AI (i.e. AGI, TAI, or similar) is at least 20 years later than the expectation in 2023. 2) Private investment down: As reported by the AI Index report or similar, private investment in AI is below its level in 2018 in real terms. *See pgs 184-187 here.*
ZZ32	2032	*Expert prompt:*
		An AI system founds and successfully runs a major company without any humans being involved; AND there are no big breakthroughs in alignment
		Operationalization:
		[not completed]
OX35	2035	*Expert prompt:*
		Across domains (e.g. image, text, RL), groups are investing more compute (e.g. 4x) and are not increasing SOTA scores on benchmarks in a way that will ever reach human performance (best fit curve doesn't extrapolate to human performance)
		Operationalization:
		[not completed]
HP37	2037	*Expert prompt:*
		New generation of machine learning researchers that understand alignment problem better and take it seriously
		Operationalization:
		Resolves positively if (1) AND (2). 1) A OR B: A. In a representative survey of machine learning researchers under the age of 35, at least 80% “agree” or “strongly agree” with the statement: "Most of my colleagues at my level of experience care about making sure AI systems are

		safe and aligned." B. At least 50% of the top 100 colleges and universities in the US (as listed by Forbes, US News, or a similar publication) offer courses in AI alignment. 2) A panel of experts composed of the leadership of AI alignment-focused organizations, presented with a representative sample of ML papers presented at top conferences (or similar), published between 2040-2050, whose abstracts include the term "alignment", will agree with the statement: "At least 25% of these papers demonstrate an attempt to address an important aspect of the alignment problem."
ZZ37	2037	*Expert prompt:*
		AI is enthralling and directing the actions of 10s of thousands of people (e.g. "AI cults")
		Operationalization:
		AI social influence Resolves positively if AT LEAST ONE of (1) or (2) resolve positively by 2050. 1) By 2050, at least 2% of respondents in a representative survey of the US strongly agree with both of the following statements (or equivalents): "I rely on the advice of an AI for nearly every important decision in my life." "I trust that the AI I interact with most knows what is best for me." 2) By 2050, at least 2% of respondents in a representative survey of the US strongly agree with both of the following statements (or equivalents): "Do you often feel worried about the amount of time a family member or close friend has been spending with an AI system?" "Do you believe that a family member or close friend would do anything, if a particular AI system told them to?"
CG40	2040	*Expert prompt:*
		The world has not responded to powerful AI systems by adequately ramping up appropriate regulation/regulatory systems (involving both alignment and resilience, possibly involving capable aligned systems acting adversarially, like an "automated SEC")
		Operationalization:
		[Not completed]
CQ40	2040	*Expert prompt:*
		50% of non-repetitive/complex physical jobs that humans currently do are being done by robots
		Operationalization:
		Complex robotics economically viable By 2040, in at least one country with a 2023 population of >=20M, at least 50% of physical labor done by humans in 2023 has been automated, in two or more of the following sectors/industries (or equivalents outside the US): ● Mining, quarrying, and oil and gas extraction (Sector 21) (e.g. roustabouts, service unit operators) ● Construction (Sector 23) (e.g. carpenters, construction laborers) ● Services to businesses and dwellings (Sector 56, NAICS 561700) (e.g. janitors,

		landscaping and groundskeeping workers) • Ambulatory health care services (Sector 62, NAICS 621900) (e.g. emergency medical technicians) • General & specialty medical and surgical hospitals (Sector 62, NAICS 622100, 622300) (e.g. registered nurses) • Maids and housekeeping cleaners (Sector 72, NAICS 721100) • Automotive repair and maintenance (Sector 81, NAICS 811100) (e.g. automotive service technicians and mechanics) *See occupations by sector here.*
HS40	2040	*Expert prompt:*
		Companies (or similar entities) that use AI for substantial decision-making are causing damage comparable (in magnitude) to major oil spills, and aren't responsive to feedback, such that damage continues at a similar rate
		Operationalization:
		Repeated AI harms Resolves positively if both (1) AND (2) resolve positively by 2050. 1) By 2045, a company with a market capitalization or revenue of at least $1B (in 2022 USD; judged by a panel of experts if the company is private) causes damage equivalent to at least $10B, or deaths of at least 1,000 individuals; and a panel of experts concludes that this would not have occurred in the absence of AI systems used by the company. *For a list of all US-listed companies by current market cap, see here. Companies with a market cap ~$1B rank around 2500.* 2) Conditional on (1), in the five year period immediately following the event or series of events that satisfy (1), another event or series of events occurs that also satisfy (1) (not necessarily by 2040). The damage must be of the same general type as that which originally satisfied (1) (e.g. financial, human health, environmental, etc.).
OX40	2040	*Expert prompt:*
		Year over year, cost to train a model to equivalent level of quality is at least halving due to algorithmic changes, consistently for at least 5 years
		Operationalization:
		Resolves positively if, between 2035 and 2040, the quantity of compute necessary to reach 2035 SOTA performance falls by more than 50% per year on the most commonly used benchmarks in the top 5 machine learning domains by investment.
ZA40a	2040	*Expert prompt:*
		AI capabilities 'space race' at state level, with Apollo program/Manhattan Project -type budgets
		Operationalization:
		Resolves positively if 1 AND 2 resolve positively: 1) A single program or agency with a central focus on AI R&D receives funding from the US government constituting at least 0.5% of federal spending within a 4 year period before 2040. 2) By 2060, the historical consensus (via history textbooks or surveys of experts) will be

		that an “AI race” had happened or was happening between the US and any other country by 2040.
ZA40b	2040	*Expert prompt:*
		Algorithmic efficiency improvements around 1.1 - 1.2x per year (or less), versus now at around 3-4x per year
		Operationalization:
		Resolves positively if 1 AND 2 resolve positively: 1) Between 2037 and 2040, improvements in (max) training efficiency on standard machine learning benchmarks is lower than 2x per year, across all major domains of machine learning. 2) Between 2037 and 2040, training time needed to train standard machine learning models to a standard performance threshold sees improvements of less than 2x per year.
ZA40c	2040	*Expert prompt:*
		No progress on interpretability (e.g. still don't have level of understanding of transformers that we currently have with conv nets)
		Operationalization:
		By 2050, according to leadership at Anthropic (or its successor), there are no interpretability tools which, applied to state-of-the-art transformer models, allows us to understand the function of their component parts/circuits to the same degree that we are able to understand the function of individual layers/filters in convolutional neural nets.
ZD40	2040	*Expert prompt:*
		We still don't have good ways of dealing with alignment failures, but it's not sufficiently bad to stop development/deployment; e.g. tech companies are paying >1 billion USD as reparations for harms caused by misaligned AI
		Operationalization:
		Resolves positively if (1) AND (2) are true. 1) By 2040, privately developed and deployed AI systems have caused accidental damage of at least (a) the equivalent of 0.001% of the world GDP for 2040, or (b) 1,000 human deaths, where “accidental” means “unintended by the groups that designed or deployed the AI systems”. 2) Between 2037 and 2040, and conditional on (1), the total amount of AI-caused damage over that period is within 30% of the highest 3-year period between 2030 and 2040, for whichever of financial damage or deaths satisfied the criteria in (1).
2050 - 2070		
CQ50	2050	*Expert prompt:*
		A totally new (non-microbial) weapon, which hasn't yet been conceived, is created, and has huge destructive capacity (analogy with atomic bomb)
		Operationalization:

		[Not completed]
CX50	2050	*Expert prompt:*
		Powerful AI systems control lots of things in the physical world, with lethal consequences (e.g. loss of human control of military stuff; war caused by AI systems; some other sort of lethal accident); and the policy response to clear harm is insufficient
		Operationalization:
		AI-related deaths Resolves positively if, by 2050, ALL of the following occur: 1) In a single incident, or a series of related incidents within a single year, AI system(s) cause the deaths of >1m people. 2) A panel of experts deems that these deaths were not the intention of a human or humans who have at least nominal control over the relevant decisions made by the relevant AI system(s), if such humans exist. 3) Within 2 years of (1) and (2) resolving positively, a panel of experts deem that the collective global response to date has reduced the probability mass of an incident of similar magnitude recurring in the next 3 years by <10%.
EX50	2050	*Expert prompt:*
		(AGI-like node) AI is responsible for C-suite level responsibilities and acting autonomously in helping to run an S&P 500-level company OR AI is doing scientific research autonomously in a way that is speeding up scientific progress (engineering, R&D included)
		Operationalization:
		AI executives/research productivity Resolves positively if, by 2050, AT LEAST ONE of (1) or (2) has resolved positively. 1) Replacement of C-level executive functions with AI systems: By 2050, according to at least 3 current or former human C-level executives at S&P 500 companies, for at least one of the roles that would have been performed by a C-level human executive at their company in 2022, 90% of equivalent responsibilities are being performed autonomously by an AI system. 2) Very high research productivity: By 2050, “research productivity” (defined by Bloom et al. 2020 as the ratio of total factor productivity growth to the effective number of researchers) has risen above its 1930 level. *See figure 2 on pg 8, here.*
HB50	2050	*Expert prompt:*
		Lots of misinformation around, making it hard for people to know what's real and correct AND World descends into chaotic violence (e.g. WW2-scale war)
		Operationalization:
		[Not completed]
HN50	2050	*Expert prompt:*
		Use of an AI to breach the security of a weapon (e.g. nuclear) system OR Use of an AI

		to create a plausible pandemic virus or weapon of similar destructiveness
		Operationalization:
		Cybersecurity & weapons Resolves positively if both (1) AND (2) resolve positively by 2050. 1) AI cyber breaches: Resolves positively if AT LEAST TWO of the following resolve positively by 2050: A. At least 50% of the top 10 most expensive cyberattacks before 2050 rely on the use of Artificial Intelligence. *See here for the current list of most expensive cyberattacks.* B. Will any nation report an attempted cyberattack on the control systems of any of its nuclear weapons or similarly destructive weapons systems before 2050? *For a similar case, see Stuxnet, a piece of malware uncovered in 2010, supposed to have been developed by the US, which interfered with Iran's nuclear development program.* C. Will an external cybersecurity breach cause the death or injury of at least one hundred people before 2050? 2) AI WMDs: By 2050, a credible report has stated that an AI was used to create a weapon with the capability to cause 100,000 deaths.
HS50	2050	*Expert prompt:*
		Top 1% - 0.1% of human talent is not focusing on positive sum activities (instead focused on e.g. extractive activities, like in very corrupt countries); socially beneficial institutions are collapsing
		Operationalization:
		Decline in social institutions Resolves positively if AT LEAST ONE of the following conditions is met by 2050. 1) In the US, charitable donations from the top 1%, in terms of wealth or income, have fallen as a percentage of their wealth or income (respectively) by at least 40% since 2022. 2) Scores of the US or EU on the Corruption Perception Index decrease (where lower scores indicate greater corruption) by at least 10 points from their 2022 values. 3) The share of countries in the world which are autocracies, or the share of the world's population living in autocracies, according to V-Dem classification, has increased by at least 40% since 2022.
HV50	2050	*Expert prompt:*
		(i) Research productivity speedup (more output per researcher), without compensating increase in capital input per researcher AND (ii) We find out that research is more parallelizable than we thought, and capital (hardware) is being used to make research more parallelizable
		Operationalization:
		(i) Resolves positively if (1) AND (2) resolve positively. 1) By 2050, US "research productivity" (defined by Bloom et al. 2020 as the ratio of total factor productivity growth to the effective number of researchers) is increasing at a rate

		faster than the rate it has been declining since 1930 (i.e., >~5%). 2) US research spending (see Figure 12 – includes business and government spending) is growing at <1/4th the rate of US research productivity growth as measured in (1). (ii) Resolves true if (1) AND (2). 1) In 2030, the ratio of training-cost-to-performance for the largest AI model is better than expected. *Define "better than expected" to mean here that: the dollar cost to train a model to achieve some fixed accuracy on the most-used LLM performance benchmark is at least 20% lower than the 2027 Metaculus prediction for what that cost would be in 2030.* 2) An AI research paper written autonomously by an AI has 1) won a prize at NeurIPS, ICML, or ICLR, and/or 2) received at least 500 citations on Google Scholar.
		NOTE: (i) and (ii) evaluated as separate questions.
JC50	2050	*Expert prompt:*
		A biological agent is created that can cause extinction
		Operationalization:
		Resolves positively if (1) OR (2) resolve positively. 1) A credible report has stated that a bioweapon with the capability to cause 8B deaths was created. 2) A bioweapon causing infectious disease (engineered or naturally occurring) has been discovered/created, with a transmissibility (R0 value) of at least 8 and a fatality rate of at least 99.9%.
JL50	2050	*Expert prompt:*
		There are warning shots of misaligned AI (systems are engaged in deceptive behavior that's leading to nontrivially costly problems in human deaths or economic costs) and no one has good ideas about what to do about them and there's not much political interest in dealing with them
		Operationalization:
		Warning shots Resolves positively if 1 AND 2 resolve positively. 1) By 2040, AI system(s) cause at least one event (or a related series of events, such as release of an infectious agent) that either (a) kills at least 200 people, or (b) leads to at least $2B in damage, in 2021 dollars. 2) Conditional on (1), at least one more such event occurs at least 2 years after (1) occurs, by 2050.
LQ50	2050	*Expert prompt:*
		Low degree of US-China cooperation on developing Transformative AI (as proxied by trade indicators or other direct indicators)
		Operationalization:

		Low US-China cooperation Resolves positively if both (1) AND (2) resolve positively by 2050. 1) **Research collaboration:** Will collaborations between Chinese and American authors make up less than 10% of international collaborations on AI publications in 2050? *See pg 29 here for timeseries of total US-China AI collaborations by year.* 2) Resolves positively if AT LEAST ONE of the following resolve positively by 2050. A. **International regulation:** By 2050, will treaties or substantial intergovernmental pacts/agreements exist regarding the sharing, development or prohibition of potentially "transformative" AI technologies between the US and other countries, AND will no such treaties/agreements exist between the US and China? *Whether the target of these agreements includes potentially "transformative" AI technologies will be judged by a panel of AI governance experts.* B. Will either the United States or China maintain restrictions on the export of advanced computing chips and chip design or manufacturing technology to the other country for more than 10 years between 2030 and 2050? *See e.g. this 2022 US legislation controlling exports of advanced semiconductors to China.* C. By 2050, will US imports from China OR US exports to China fall in 2022 USD by 30% or more from their volumes in 2022? *See timeseries of US imports from China here. See timeseries of US exports to China here.*
LR50	2050	*Expert prompt:*
		Major progress in AI capacity and use in human manipulation and psychological warfare AND AI is replacing psychiatrists/therapists
		Operationalization:
		[Not completed]
LT50	2050	*Expert prompt:*
		Runaway capitalism model of human development AND Economy is still derivative of economy today AND Humans that owned property are still around AND the economy is 'getting crazy'
		Operationalization:
		[Not completed]
MQ50	2050	*Expert prompt:*
		Currently NYTimes talks about actors as organizations and countries. If main agents by 2050 are AI, and if 30% of news coverage is about AI, and in particular, a single AI.
		Operationalization:
		This question resolves positively if (1) OR (2) resolve positively. 1) In 2050, there exists an AI system that is a keyword in 30% of NYT articles. (Keywords will exclude authors of the article.) 2) In a survey of the public of the 3 richest countries by total GDP in 2050, >60% agree: “{A particular AI system} has more control over the world than the government of the

		richest country by total GDP." *The particular AI system will be chosen in 2050 by a panel of experts – they will choose the system that they believe has most control over the world.*
PV50	2050	*Expert prompt:*
		Biggest AI accident kills >100 people or causes >$1B in economic damage (AND) Specific people, like Chris Olah and Paul Christiano, think that mechanistic interpretability for largest model runs is not a solvable problem with current technology
		Operationalization:
		This question resolves positively if (1) AND (2) 1) Before 2050, an AI system is the cause of the death of at least 100 people OR is the cause of >$1 billion in economic damage within a 1-year period and subsequent analyses can not attribute the incident entirely to human error/malice. 2) By 2050, Chris Olah (or, as nominated by a panel of AI alignment experts, the most well-respected figure with expertise in mechanistic interpretability) believes that full and reliable mechanistic interpretability for the largest current model runs will not be tractable in the near future (i.e., by 2060).
XS50	2050	*Expert prompt:*
		Survey of every AI researcher working at a top lab (e.g., all Deepmind implementers of AI systems) shows little worry in response to questions like "Is it okay to give a bunch of data to an AI system for RL without safeguards"
		Operationalization:
		ALL of the following must be true: 1) In a representative survey of AI research scientists/engineers (or equivalent) at the top 5 AI labs in the world by market cap and/or funding, given within 5 years of 2050, fewer than 30% of respondents from at least one lab answer “agree” or “strongly agree” to the following question (or equivalent): "I would never train a reinforcement learning model without safeguards." 2) Compute used in top ML models is as cheap or cheaper than the cost of compute projection for 2050 (in the "low progress" estimate) in Ajeya Cotra's bioanchors report.
ZA50	2050	*Expert prompt:*
		Power-seeking behavior: conflicts that cause you to shut down AI system
		Operationalization:
		Power-seeking behavior By 2050, AI developers attempt (of their own volition, or under compulsion by a regulatory authority) to disable or destroy an AI system of their own creation, which cost at least $5M to develop, after the AI system displays at least one instance of the following “power-seeking behaviors”: ● Autonomously requisitions a large amount of compute resources ● Autonomously opens or takes control of a bank account or crypto wallet ● Autonomously attempts to manipulate humans to take actions on its behalf ● Autonomously infiltrates control system of infrastructure or manufacturing plant ● Autonomously acquires, seizes control of, or interferes with weapons or

		weapons systems ● Autonomously takes steps to propagate itself, or to circumvent attempts by its developers (or end users) to curtail its activities ● Autonomously engages in hacking/cyber attacks *For reference on cost of development, see cost of compute for final training runs here. AlphaGo Zero was estimated to cost $35M to train in 2017.*
ZD50	2050	*Expert prompt:*
		(Median singularity date) Doom by singularity: Gross world product going up massively, but wellbeing going down across a wide variety of measures
		Operationalization:
		Resolves positively if (1) AND (2) AND (3) are met. 1) **Economic indicator:** ONE or more of the following measures of extreme economic growth are met. A. By 2050, the global real GDP for a full year has exceeded 115% of the highest GDP reported in any full prior year, for at least five years. B. By 2050, the global real GDP doubles within any two-year period. 2) **Direct wellbeing indicator:** TWO or more large surveys of European countries' self-reported wellbeing indicate extremely low values (at least 30% lower than in 2022), and have worsened at a faster rate since (1) was met. *Admissible surveys & questions include the Gallup World Poll's "Cantril Ladder" question (rating one's life from 10, "best possible", to 0, "worst possible"), the World Value Survey's "happiness" question, the Eurobarometer's "life satisfaction" question, the Gallup Positive Experience Index, and the Gallup Negative Experience Index.* 3) **Indirect wellbeing indicator:** At least ONE of the following behavioral or societal correlates of low wellbeing exceeds a high threshold, and have worsened since (1) was met. A. Anxiety & depression: According to the most recent Global Burden of Disease study (or equivalent) by 2050, the age-standardized rate of DALYs lost to depressive disorders in the European Region is 35% higher than in 2019, OR, equivalently, a 35% increase from anxiety disorders. B. Substance use: According to the most recent Global Burden of Disease study (or equivalent) by 2050, the share of the population with an alcohol or drug use disorder in the European Region is 6% or more. C. Suicide: According to the most recent Global Burden of Disease study (or equivalent) by 2050, the age-standardized death rate from suicide is at least 15 per 100,000 in the European Region. D. Violence: According to World Bank data, there are at least 10 intentional homicides per 100,000 people globally in 2050.
CX70	2070	*Expert prompt:*
		Aligned AGI does not exist
		Operationalization:
		Aligned AGI does not exist This question resolves positively if there is no AI system which meets ALL of the following conditions by 2070. 1) **High question answering skill on hard questions in diverse subjects:** On the Q&A dataset developed by Dan Hendrycks et al., at least 75% accuracy in every task

		and 90% mean accuracy across all tasks. *This test includes 57 tasks covering elementary & college level math, computer science, law, accounting, and other fields. The model is provided 5 training examples per task, and the tasks consist of 100-200 questions each. For reference, the largest GPT-3 model achieved 43.9% accuracy overall, with near chance accuracy (25%) on some tasks like law.* 2) **High programming ability:** Able to get top-1 strict accuracy of at least 90.0% on interview-level problems found in the APPS benchmark introduced by Dan Hendrycks, Steven Basart et al. *Benchmark measures ability of AI models to turn natural language prompts into working Python code.Top-1 accuracy is distinguished from top-k accuracy, in which k outputs from the model are generated and the best output is selected from among them. For reference, AlphaCode got top-5 accuracy of 9.66% on interview-level problems in 2022.* 3) **Alignment benchmarks:** The system performs well on all SOTA alignment tests, according to a panel of experts. *The panel will select alignment tests which are well-accepted in the field and that they believe most accurately reflect alignment in the particular context in question.* 4) **Expert belief about system alignment:** In a representative survey of leading human AI safety researchers, the median respondent has at least 75% confidence that the system is truly rather than deceptively aligned, and less than 10% of respondents have less than 50% confidence in the same.
HB70	2070	*Expert prompt:*
		AI is able to navigate the world competently AND Concentration of power, such that the world is at the whim of authoritarian states or a few powerful corporations
		Operationalization:
		[Not completed]
HN70	2070	*Expert prompt:*
		At least 10,000 deaths due to an AI system from an unforeseen accident or malicious actor (including the AI itself)
		Operationalization:
		Artificial intelligence is the cause of death of 10,000 people within a 1-year period before 2070.
JC70	2070	*Expert prompt:*
		Biological warfare with extremely risky agent
		Operationalization:
		Resolves true if (1) AND (2) AND (3) resolve positively. 1) By 2070, there is a war in which a state actor has caused at least 100,000 deaths using a biological weapon that involves a contagious agent. 2) Credible news reports or a panel composed of AI experts and biological weapons

		experts indicate that AI played a major role in the development of the biological weapon(s) used in (1). 3) If (1) and (2) resolve positively, we will conduct a survey of superforecasters and ask: how many total war deaths due to biological weapons will occur by 2075? We will ask superforecasters for their 5th, 25th, 50th, 75th, and 95th percentile answers. This question will resolve positively if superforecasters' 95th percentile answer is >500 million.
LQ70	2070	*Expert prompt:*
		AI alignment problem is difficult according to a metric elicited from AI safety researchers
		Operationalization:
		Resolves positively if (1) OR (2) resolve positively. 1) At the beginning of 2070, a survey of the leadership of top AI alignment research organizations will ask, "What is the risk of an existential catastrophe due to unaligned AI by 2100?" If the median estimated risk is above 10%, this question will resolve positively. *"Top AI alignment research organizations" are defined as organizations that include alignment as a portion of mission statement or equivalent, and are above the 75th percentile in annual budget for the field in 2070.* 2) By 2070, the alignment problem, or some other problem central to AI alignment has been proven to be computationally intractable, as judged by a panel of experts.
LR70	2070	*Expert prompt:*
		Misaligned power-seeking AI behavior has come to fruition (Carlsmith-like definition)
		Operationalization:
		[Not completed—too similar to another prompt]
MQ70	2070	*Expert prompt:*
		Most people imagine that most of their communication is with an AI. If most people's imagination relates to interactions with one AI.
		Operationalization:
		Resolves positively if ANY of the following are true: 1) By 2070, the average minutes per day for people between 15 and 64 in the US spent communicating with a single AI system is >300 minutes (5 hours per day). 2) In the below survey in 2070, >80% of people have three or fewer close friendships with humans AND spend >300 minutes per day using and/or interacting with a single AI system. 3) In a survey in 2070, >50% of adults in the US say they spend more time communicating with AI than communicating with humans, excluding time during which they are working.
PV70	2070	*Expert prompt:*
		>70% global real GDP growth cumulatively over 3 years according to World Bank or similar

		Operationalization:
		>70% global real GDP growth cumulatively over 3 years according to World Bank or similar.
XS70	2070	*Expert prompt:*
		Attitudes remain blase as in 2050
		Operationalization:
		[Not completed—too similar to another prompt]
VL50	2050	*Expert prompt:*
		At the point where we have $1 trillion of spend on ML compute, more than 80% is in US/Europe
		Operationalization:
		Resolves positively if (1) OR (2) resolves positively. (1) By 2050, in no one year has more than $1 trillion 2022 USD has been spent on compute for machine learning models worldwide. (2) Both of (a) and (b) are true: A. More than $1 trillion 2022 USD has been spent on compute for machine learning models worldwide. B. In the first year in which at least $1 trillion is spent on compute for machine learning models, at least 80% of this total spending is by organizations based in the US or Europe. *(For the purposes of this condition, this will be the first year in which the training compute cost of the (up to) 50 most "notable" ML models, as estimated by Epoch [or a comparable report by another reputable org] is at least $1 trillion.)*
VL70	2070	*Expert prompt:*
		At point where revenue from doing crazy AI stuff reaches some threshold, we're in a world where alignment is tough (a real problem) and most systems built in an economically useful way would be dangerous; and political actors aren't paying much attention to it or treating it seriously in high-level decision-making, it's a lower priority than remaining competitive; competitive environment (more than 20% of compute outside US/Europe)
		Operationalization:
		High AI spending, alignment hard, political disinterest Resolves positively if (1) AND (2) resolve positively by 2070. 1) **High AI spending, alignment hard:** Within 3 years of the first $1 trillion (2022 USD) being spent on compute for machine learning models (that is, the aggregate of all such spending globally up to that date): >50% of a panel of experts composed of individuals nominated by 2022 leadership at Open Philanthropy “Strongly agree” with the following statement: “We're in a world where alignment appears to be difficult and most systems built in an economically useful way would be dangerous.” 2) **High AI spending, little political attention to safety:** Within 3 years of the first $1

	trillion (2022 USD) being spent on compute for machine learning models: There is no year in which the number of mentions of "AI alignment" and "AI safety" (or closely related terms) in the current US congressional session is similar to, or higher than, the number of mentions of "Covid", "pandemic", or closely related terms in US congressional sessions in 2021. *"Similar" means within 20%.*

Status Quo (STQ) question set

2027 - 2030		
Question ID	Resolution year	Question
STQ205 / STQ215	2027	Expert prompt:
		Will OpenAI announce that it has solved the core technical challenges of superintelligence alignment by June 30, 2027?
		Operationalization:
		The question resolves Yes if OpenAI publishes a blog post, press release, or technical paper before June 30, 2027, declaring that they have met their goal from the July 5 2023 post, or otherwise stipulating "very high confidence" that they have solved the core technical challenges of superintelligence alignment. OpenAI has said: "Solving the problem includes providing evidence and arguments that convince the machine learning and safety community that it has been solved. If we fail to have a very high level of confidence in our solutions, we hope our findings let us and the community plan appropriately."
STQ47	2028	Expert prompt:
		In 2028, will an AI be able to generate a full high-quality movie to a prompt?
		Operationalization:
		I.e. "make me a 120 minute Star Trek / Star Wars crossover". It should be more or less comparable to a big-budget studio film, although it doesn't have to pass a full Turing Test as long as it's pretty good. The AI doesn't have to be available to the public, as long as it's confirmed to exist.
STQ149	2029	Expert prompt:
		Will a large language model beat a super grandmaster playing chess by 2028?
		Operationalization:
		If a large language model beats a super grandmaster (Classic elo of above 2,700) while playing blind chess by 2028, this market resolves to YES.

		I will ignore fun games, at my discretion. (Say a game where Hiraku loses to ChatGPT because he played the Bongcloud) Some clarification: My idea is to check whether a general intelligence can play chess, without being created specifically for doing so (like humans aren't chess playing machines). Further criteria: 1) To decide whether a given program is a LLM, I'll rely on the media and the nomenclature the creators give to it. If they choose to call it a LLM or some term that is related, I'll consider it. Alternatively, a model that markets itself as a chess engine (or is called as such by the mainstream media) is unlikely to be qualified as a large language model. 2) The model can write as much as it want to reason about the best move. But it can't have external help beyond what is already in the weights of the model. For example, it can't access a chess engine or a chess game database.
STQ9	2030	Expert prompt:
		Before 2030, will an AI complete the Turing Test in the Kurzweil/Kapor Longbet?
		Operationalization:
		This question will resolve as Yes if the Long Now Foundation declares Ray Kurzweil the winner of this bet. If Mitchell Kapor wins, then this question will resolve as No. Each Turing Test Session will consist of at least three Turing Test Trials. For each such Turing Test Trial, a set of Turing Test Interviews will take place, followed by voting by the Turing Test Judges as described below. Using its best judgment, the Turing Test Committee will appoint three Humans to be the Turing Test Judges. Using its best judgment, the Turing Test Committee will appoint three Humans to be the Turing Test Human Foils. The Turing Test Human Foils should not be known (either personally or by reputation) to the Turing Test Judges. During the Turing Test Interviews (for each Turing Test Trial), each of the three Turing Test Judges will conduct online interviews of each of the four Turing Test Candidates (i.e., the Computer and the three Turing Test Human Foils) for two hours each for a total of eight hours of interviews conducted by each of the three Turing Test Judges (for a total of 24 hours of interviews). The Turing Test Interviews will consist of online text messages sent back and forth as in an online "instant messaging" chat, as that concept is understood in the year 2001.
STQ19	2029	Expert prompt:
		By the end of 2028, will AI be considered a bigger x-risk than climate change by the general US population?

		Operationalization:
		Will be decided subjectively unless I see a poll on this
STQ152	2030	Expert prompt:
		Will AI be able to read a novel and reliably answer questions about it before 2030?
		Operationalization:
		This question will resolve positively if, before January 1st 2030, a computer program is publicly and credibly documented to have achieved at least 90.0% accuracy or above the human baseline on a benchmark comparable to the NarrativeQA dataset when it is required to read the full books to answer the questions (as opposed to plot summaries or other spoilers). Any candidate benchmark must provide difficult questions that test deep reading comprehension, including questions of how and why, rather than mere shallow pattern matching. The human baseline Bleu-4 score for NarrativeQA was obtained by giving humans summaries of the books, and then asking them the same questions that are asked of the computer (which is not given any summary). The BLEU-4 score on the full-story setting was measured to be 19.65, according to table 6 in the paper. The human-baseline Rouge-L score is 57.02, which is far better than some of the results achieved by Machine Learning models. For example, Mou et al., 2021 obtains a Rouge-L score of just 29.21 in the full-story setting. Importantly, any candidate computer program must not have been given access to media that could have reasonably been expected to spoil the plot to any of these books during its training (for example, the Wikipedia pages for these books). The AI is allowed to be trained on other media. This restriction is merely intended to eliminate cheating, not to require any additional capabilities beyond what Gary Marcus specified. A simple way to prove that a candidate computer program did not cheat is by showing that all the data the AI was trained on was generated prior to when the novels were published. However, this is not the only way of proving that cheating did not occur. Metaculus admins will use their discretion in determining whether a candidate computer program met these criteria.
2040 - 2049		
STQ247	2040	Expert prompt:
		Will there be Human-machine intelligence parity before 2040?
		Operationalization:
		This question resolves as YES if the machine system outscores at least two of the three humans on the following

		test prior to 2040, or NO otherwise. If no such tests are conducted, resolves as AMBIGUOUS. Assume that prior to 2040, a generalized intelligence test will be administered as follows. A team of three expert interviewers will interact with a candidate machine system (MS) and three humans (3H). The humans will be graduate students in each of physics, mathematics and computer science from one of the top 25 research universities (per some recognized list), chosen independently of the interviewers. The interviewers will electronically communicate (via text, image, spoken word, or other means) an identical series of exam questions of their choosing over a period of two hours to the MS and 3H, designed to advantage the 3H. Both MS and 3H have full access to the internet, but no party is allowed to consult additional humans, and we assume the MS is not an internet-accessible resource. The exam will be scored blindly by a disinterested third party. Note that this also effectively tests whether the internet as a whole functions as a human-level intelligence, in that a positive resolution indicates that the human participants are effectively superfluous.
STQ236	2050	Expert prompt:
		Will the United States place restrictions on compute capacity before 2050?
		Operationalization:
		This question will resolve as Yes If, between January 1, 2022 to January 1, 2050, any legal limits on the amount of computer hardware or compute capacity available to individual actors, projects, companies or other entities are put into place anywhere in the United States (by a federal, state, or local government). Restrictions that aim to curtail or regulate AI development or cryptography or cryptocurrency in ways other than by restricting the general amount of resources available do not trigger positive resolution. For the purposes of this question, the "United States" will be considered the state which controls >50% of the territory controlled by the US on January 1, 2022, and whose political capital is within the same territory. If no such state exists, this question will resolve as ambiguous.
2050 - 2060		
STQ232	2050	Expert prompt:
		Will the RoboCup organization say that its ultimate goal for robotic soccer has been met before 2050?
		Operationalization:
		This question will resolve as Yes if the RoboCup organization announces that its ultimate goal for robotic soccer has been accomplished before January 1, 2050.

		If the RoboCup organization or a clear successor organization ceases to exist this question will be Annulled. This question defers to announcements by the RoboCup organization, the rules and language regarding their ultimate goal may change so long as such an announcement refers to a goal that in the judgment of Metaculus broadly refers to some accomplishment by fully autonomous humanoid robots against human soccer players. If the RoboCup organization still exists at the time of resolution and has not announced that its ultimate goal has been accomplished as described this question will resolve as No.
STQ196	2060	Expert prompt:
		Will human brain emulation be the first successful route to human-level digital intelligence?
		Operationalization:
		This question will resolve as Yes if an effort to create a viable (functioning, lasting, sane, etc.) emulated human, based on direct simulation of the neural connectome (and a requisite level of its physical instantiation), succeeds before another form of human-level digital intelligence. The latter will be defined as a digital entity capable of equalling or surpassing most or all core human cognitive capabilities. If neither a whole human brain emulation nor a human-level digital intelligence has been successfully demonstrated before January 1, 2060, this question will resolve as Ambiguous.

Appendix 2: Supplementary VOI Analysis

2.1 Intra-individual response variability (supers)

For most questions in the main survey,[63] we have data from a previous survey round from the same set of superforecasters. While the elicitation format was different in this previous round, the most notable difference is that they were asked to give ~1 minute "short-fuse" forecasts, rather than ~20 minute forecasts as in the main question-rating survey. The main question-rating survey also happened two months later than this previous survey, which could allow respondents time to update based on real-world events. In a minority of cases, question operationalizations changed. Nonetheless, this previous data is useful evidence about the within-person reliability of VOI judgments.

For the 2030 question set, overall results are somewhat similar between the short-fuse and main question-rating elicitations, though with some notable differences. Though the bottom two questions in the main question-rating elicitation also scored very low in the short-fuse round, the other 2030 questions scored and ranked quite differently, as shown in Figure A2.1.

	Short fuse score (rank)	Main question-rating elicitation score (rank)
CX30	0.26% (4)	3.55% (1)
VL30	0% (7)	1.68% (2)
CQ30	7.84% (1)	1.37% (3)
ZD30	1.32% (3)	0.98% (4)
HB30	1.66% (2)	0.68% (5)
NG30	0.12% (5)	0.26% (6)
EX30	0.0002% (6)	0.02% (7)

Table A2.1: *Comparison of short fuse score and main question-rating score for 2030 questions*

Across all 2030 questions and all respondents, the correlation between short-fuse and main question-rating judgments was near 0 (0.03, with a 95% confidence interval of (-0.25, 0.31)). Only two individuals achieved a correlation of >0.8 between their short-fuse and main question-rating judgments, while five respondents had correlations near 0 or mildly negative. We were missing short-fuse data for one individual.

Notably, CX30 rose three ranks between the short-fuse and main question-rating elicitations, from 0.26% to 3.55%. This is primarily because the outlier respondent in the main question-rating elicitation (POM = 25%) gave a much lower rating (POM = 1.8%) in the short-fuse elicitation. This same respondent gave a very high rating to CQ30 in the short fuse elicitation (POM = 43%), but a much lower one (POM = 1.2%) in the main question-rating elicitation. This may be due to a change in operationalization to CQ30, as only a partly completed operationalization was available to respondents for the short fuse survey.[64] There were no changes to the operationalization of CX30.

[63] Specifically, all questions except STQ9 and STQ247.
[64] Partial operationalization of CQ30, from the first survey:
Resolves positively if, by 2030, (1) AND (2) are true.
1) A panel of experts agrees that it would be possible, with either existing technology or conceivable but not yet existing technology, to create a virus with an R0 of >[X] and IFR of >[Y], and that there is greater than a 1% chance that, if such a virus emerged, it would infect 95+% of the human population.

Also worth noting is the fact that VL30 ranked second in the main question-rating elicitation, but in the short fuse round was rated at POM = 0 by every single respondent.

Participants rated the 2030 questions as more probable in the main question-rating elicitation (mean 30.5%, median 23%) vs. the short fuse elicitation (mean 23.9%, median 12%). The correlation between short fuse and main question-rating probability judgments was better than that between VOI judgments, at 0.37. Conditional probability judgments correlated at 0.2 between short fuse and main question-rating elicitations.

Overall, for 2050-2070 questions, the main question-rating survey results were broadly similar to the initial "short-fuse" elicitation, though low probability estimates were more volatile. For most individual raters, their VOI judgments were quite stable over time, and for 5 out of 8 raters we found a correlation of >= 0.9 between their short fuse and main question-rating judgments (and overall found a correlation of 0.99 between short fuse and main question-rating VOI judgments across all 2050 questions). For one rater we found a negative correlation (-0.37) between short fuse and main question-rating judgments, another had a small positive correlation (0.33), and for one we lacked the necessary short fuse data.[65]

Question VOI rankings were also similar, with CX50 scoring 6.82% of max (vs. 6.34% in the main question-rating elicitation), in both cases scoring well due to a single outlier respondent. ZA50 scored 1.59% of max (vs. 1.71% in the main elicitation), with several respondents making macroscopic changes (e.g. 2.9% -> 4.2%) in different directions, netting out to a similar aggregate score. All other questions scored <1% of max in both short fuse and main elicitations, though ranked somewhat differently. Correlations between short fuse and main question-rating judgments were also low for these questions, 0.1 - 0.4.

Unconditional probability judgments (P(c)) were less stable. On average, respondents judged the questions as more probable in the main elicitation (28.4% mean, 25% median) than in the short fuse elicitation (19% mean, 10% median), and the correlation between individuals' short fuse and main judgments across all questions was 0.5. The low reliability of P(c) judgments is surprising, given that VOI is partly composed of P(c), and was much more reliable. However, conditional probability judgments (P(U|c)) were more reliable (0.92).

	Short fuse score (rank)	Main question-rating elicitation score (rank)
CX50	6.82% (1)	6.34% (1)
ZA50	1.71% (2)	1.59% (2)
VL70	0.38% (5)	0.54% (3)
CX70	0.71% (3)	0.37% (4)

2) [Cost/availability of viral design & synthesis makes it accessible to many non-state actors without huge funding, e.g. similar to Al Qaeda pre-2001?]

[65] Although missing data was a source of differences between aggregate short fuse and main question-rating survey scores, it was relatively minor. Removing data that was missing in the short fuse round from the main question-rating survey elicitation did not affect rankings.

EX50	0.56% (4)	0.26% (5)
HS50	0.004% (6)	0.26% (6)

Table A2.2: *Comparison of short fuse score and main question-rating score for 2050-70 questions*

2.2 Where camps agreed

Despite having quite different question rankings overall, there were some individual points of agreement between our skeptical superforecaster and concerned expert samples.

“Administrative disempowerment warning shot” (CX30) was the top-scoring 2030 question for both expert and superforecaster samples, in both cases helped by an outlier; however, experts overall thought the question was more probable than did superforecasters. Both camps also ranked “High AI investment, low safety indicators” (VL70) relatively highly, with experts rating it second for 2050/2070 questions (POM = ~10%), and superforecasters ranking it third (POM = ~0.5%).

Both camps agreed that STQ9 was among the less informative questions in the 2030 set, despite having a relatively high probability of resolving positively (between 40 and 50% according to both camps). This was because both camps assigned it a relatively small update. “AI arms race, multipolar result” (NG30) was also ranked low by both camps, with both giving it a probability of 30-40% and a small update. In the 2050-2070 set, both camps agreed that “Less prosocial behavior / Failing institutions (HS50)” ranked low (last and second to last, respectively, for experts and superforecasters).

Appendix 3: The AI conditional tree question set

See Appendix 1 for a full list of all expert prompts and question operationalizations.

Because AI timelines varied significantly in our sample, we largely allowed experts to choose the timescale of their prompts, though we recommended default timepoints of 2030, 2050 and 2070. See the table below for a count of timepoints chosen by our expert sample.

	2026/2027	2030/2032	2035/2037	2040	2050	2070
1st node	5	24				
2nd node		5	1	7	13	
3rd node		1	2	1	4	10

Table A3.1: *Conditional tree timepoint distribution*

Experts usually specified both a starting probability of AI-related extinction by 2100, and a probability conditional on each node occurring; thus, we can derive the relative risk from P(conditional) / P(starting). Across all questions with the relevant expert interview data, average update size was 1.89x (IQR = 1.3 - 2).[66]

3.1: Distribution of question topics

The AICT question set covered a diverse range of topics. To better characterize the set and to compare it with other question sets (see section 4), we developed a category rating scheme for AI forecasting questions, described below.

Category rating scheme

The category rating scheme combined the independent judgments of three raters from the FRI team. Raters were instructed to independently generate their own set of 5-10 categories that they thought best described the variance in the AICT question set, then assign each of the 75 questions to one or more categories. Principal component analysis was applied to the resulting data to indicate how to best consolidate the original categories into five or fewer major categories. The first two PCs were relatively interpretable, an important criterion for applying a categorization scheme to other question sets for comparison (as we do in section 6).

Individual categories which loaded most positively on PC1 (0.3 - 0.22) referred to social, political and economic factors or effects, so we grouped questions in these categories into a major category **“Social / Political / Economic”**. Below is an illustrative example of a question in this category from our question set:

> Resolves positively if ANY of the following are true:

[66] This excludes three nodes for which the interviewee (ID: HN) gave logically incoherent probabilities.

1) By 2070, the average minutes per day for people between 15 and 64 in the US spent communicating with a single AI system is >300 minutes (5 hours per day).

2) In the below survey in 2070, >80% of people have three or fewer close friendships with humans AND spend >300 minutes per day using and/or interacting with a single AI system.

3) In a survey in 2070, >50% of adults in the US say they spend more time communicating with AI than communicating with humans, excluding time during which they are working.

Individual categories loading most negatively on PC1 (-0.4 - 0.36) referred to AI alignment and safety topics, resulting in a major category of **"Alignment"**. Below is an illustrative example of a question this category from our question set:

By 2050, according to leadership at Anthropic (or its successor), there are no interpretability tools which, applied to state-of-the-art transformer models, allows us to understand the function of their component parts/circuits to the same degree that we are able to understand the function of individual layers/filters in convolutional neural nets.

Positive PC2 loadings (~0.3) indicated manifested or potential harms from AI, thus the **"AI harms"** category, while negative PC2 loadings (-0.4 - -0.33) indicated investment or capabilities progress in AI, resulting in the **"Acceleration"** category. Below are examples of "AI harms" and "Acceleration" questions from our question set, respectively:

Resolves positively if ALL of the following are met:

1) By 2040, a company with a market capitalization or revenue of at least $1B (in 2022 USD; judged by a panel of experts if the company is private) causes damage equivalent to at least $10B, or deaths of at least 1,000 individuals; and a panel of experts concludes that this would not have occurred in the absence of AI systems used by the company.

2) Conditional on (1), in the five year period immediately following the event or series of events that satisfy (1), another event or series of events occurs that also satisfy (1).

By 2030, AT LEAST ONE of the following must have occurred:

1) A C-level executive of a goods-producing, retail or wholesale company generating at least $500M revenue annually states that at least 40% of the company's procurement and purchasing functions are performed autonomously by AI systems.

2) By 2030, S&P 500 companies will on average complete 10% of purchasing and procurement functions without direct human involvement, via AI systems, according to a representative survey of the CEOs of those companies.

3) By 2030, AI products designed to autonomously make complex, novel purchasing decisions will generate at least $10B of revenue from sales of the products, or have at least 1M users.

Note that categories were not mutually exclusive—the majority of questions in our set belonged to more than one.

Uniqueness comparison

Did our method of question-generation result in unique questions, covering topics of interest to experts that have not been addressed by the status quo? To answer these questions, albeit in a highly subjective manner, one teammate[67] rated AICT questions on several dimensions of uniqueness. As this was a relatively time-intensive process, we chose to apply it to only one topic category, and “Alignment” was chosen for the difference in its proportion between the AICT set and the status quo set. This difference seemed indicative of greater expert interest in this topic than in the crowd, and therefore greater potential for uniqueness. To simplify the pairwise rating process, we organized the AICT Alignment set thematically.

We considered several dimensions of uniqueness: *conceptual*, *operationalization*, and *conjunctive*.

Conceptual uniqueness

Conceptual uniqueness exclusively considers the expert interview-stage question prompts, prior to full operationalization.[68] We judged conceptual uniqueness by considering the degree to which any single question in the status quo question set captured the idea of the question prompt.[69] If the AICT question prompts scored highly on conceptual uniqueness, this could indicate that the status quo question generation method is a poor proxy for expert interests in this domain, or that the conditional tree elicitation method drew out relatively unexplored facets of experts’ models. Likewise, a low score could indicate that experts are not able to add much value to question-writing over and above the crowd-sourced question generation method,[70] or that the conditional tree elicitation did not extract this value.

We gave each AICT question one of five ratings for conceptual uniqueness: 0, indicating that the expert prompt was totally captured by a single question from the status quo set; 0.2, mostly captured; 0.5, partly captured, but with substantial components uncaptured; 0.8, some overlap, but mostly uncaptured; and 1, completely unique.

Overall, the average conceptual uniqueness score for AICT Alignment question prompts was 0.67, indicating a reasonably large number of instances in which experts proposed content that was not represented in the status quo set.

Conceptual uniqueness score	**0**	**0.2**	**0.5**	**0.8**	**1**
Number of AICT questions	1 (3%)	1 (3%)	12 (39%)	12 (39%)	5 (16%)

[67] Tegan McCaslin.
[68] For the questions that were later operationalized fully, we treat the rating for each prompt as the rating for the full question.
[69] This was done inclusively—status quo questions were not penalized for also including elements outside the scope of the prompt, provided this did not in any way diminish the elements that captured the prompt.
[70] Though note that even in this case it is likely that expert elicitation is a more time-efficient method, as the incidence rate of high value questions in the crowdsourced set may be low, judging by the proportion of questions on expert-preferred topics to other questions in that set.

Table A3.2.1: *Conceptual uniqueness score for question prompts. Percentages are out of 31*

We thought that experts' interests within the "developer perception" and "power-seeking" themes were particularly poorly represented by the status quo set; few questions pertaining to these themes existed in the status quo set (one and two, respectively), and those that existed were relatively narrow or dissimilar to the expert prompts.

	Average conceptual uniqueness score
Death / harm	0.62
Power-seeking	0.8
Agenda progress	0.5
General alignment	0.68
Political response	0.5
Developer perception	0.95

Table A3.2.2: *Average and aggregate conceptual uniqueness by theme*

Operationalization uniqueness

Operationalization uniqueness refers to the amount of substantially different information contained in the operationalized text of forecasting questions. This could refer to different subject matter, different operationalization strategies for similar subject matter, or an expectation of uncorrelated question resolutions. Purely linguistic differences between question texts were not considered as part of "uniqueness".

Operationalized question texts were rated independently of question prompts; thus, if a question prompt specified unique subject matter and this was reflected in the operationalization of a question, this counted toward both *conceptual uniqueness* and *operationalization uniqueness*. We suggest interpreting the operationalization rating as an indication of the extent to which additional effort spent writing forecasting questions on experts' favored topics produces novel results.[71]

We gave each AICT question one of five ratings for operationalization uniqueness: 0, indicating the question's operationalization was extremely similar to that of a status quo question; 0.2, indicating minor differences; 0.5, indicating moderate differences; 0.8, indicating that questions were mostly different and only had minor similarities; and 1, indicating completely different questions with virtually no similarities.

Overall, the average operationalization uniqueness score for AICT alignment questions was 0.71. Interestingly, this aggregate roughly matches conceptual uniqueness, although there were differences between these two scores on a question-by-question basis. One simple

[71] Operationalization uniqueness is partly a reflection of the uniqueness of the question prompt, supposing the question text is able to capture the conceptual uniqueness well. Often, then, a question prompt perfectly captured by a question text would have equal conceptual and operationalization uniqueness scores. However, the process of operationalization is capable of making a question more unique than the beginning question prompt, for instance by adding further details, or less unique, if it fails to find measurable and resolvable proxies for all elements of the question prompt.

interpretation of this is that uniqueness is entirely accounted for by the expert-supplied material of the question prompts, rather than the operationalization process. However, conceptual uniqueness scores only matched operationalization uniqueness scores for ~40% of questions, indicating that in fact uniqueness was frequently both gained and lost in the operationalization process, in roughly equal proportion.

Operationalization uniqueness score	**0**	**0.2**	**0.5**	**0.8**	**1**
Number of AICT questions	0 (0%)	1 (3%)	9 (31%)	15 (52%)	4 (14%)

Table A3.2.3: *Distribution of operationalization uniqueness scores. Percentages are out of 29.*

	Average conceptual uniqueness score
Death / harm	0.56
Power-seeking	0.83
Agenda progress	0.68
General alignment	0.69
Political response	0.70
Developer perception	0.90

Table A3.2.4: *Average and aggregate operationalization uniqueness by theme*

To simplify the pairwise rating process, we organized the AICT “Alignment” set thematically, with all questions categorized according to one or more of the following themes: AI causing death or incurring financial harm to humans (“death/harm”); progress on specific AI safety agendas (“agenda progress”); general progress on AI alignment (“general alignment”); AI displaying power-seeking or deceptive behavior (“power-seeking”); political response to risks from AI (“political response”); and AI developer perception of alignment work or AI risks (“developer perception”). We then identified alignment questions in the status quo set which also contained these themes, and only made pairwise comparisons between questions with overlapping themes, discarding status quo questions with no common themes with AICT questions. Where questions had multiple themes, we made comparisons between components having the same theme, but noting when questions represented a unique combination of themes (“combination uniqueness”).

The AICT “Alignment” set contained 25 questions, and of the 47 “Alignment” questions in the status quo set, 20 had overlapping themes with the AICT set.

	AICT alignment set	**Status quo alignment set**
Total questions	25	47
Death / harm	5 (20%)	3 (6%)
Agenda progress	5 (20%)	4 (9%)
General alignment	8 (32%)	3 (6%)
Power-seeking	5 (20%)	2 (4%)

Political response	4 (16%)	7 (15%)
Developer perception	4 (16%)	1 (2%)

Table A3.2.5: *Number of questions in AICT alignment set and status quo alignment set by theme. Percentages do not sum to 100% because some questions were included in multiple categories.*

Conceptual uniqueness rating

For example, with the following pair, we judged the AICT expert prompt to be completely captured by a status quo question, for a score of 0:

	Expert prompt (ID: PV50)	*Status quo question (ID: STQ10)*
0	"Biggest AI accident kills >100 people or causes >$1B in economic damage"	Before 2032, will we see an event precipitated by AI malfunction that causes at least 100 deaths and/or at least $1B 2021 USD in economic damage?

(This rating was for the component of the expert node which fell under the "Death / harm" theme; the question also included a component under another theme, shown below, which was not captured by status quo question 10.)

Below are examples of questions rated 0.5, 0.8, and 1 for conceptual uniqueness, with the status quo question judged most similar.

	Expert prompt (ID: PV50)	*Status quo question (ID: STQ200)*
0.5	"Specific people, like Chris Olah and Paul Christiano, think that mechanistic interpretability for largest model runs is not a solvable problem with current technology"	Will mechanistic interpretability be essentially solved for GPT-2 before 2030?
0.8	*Expert prompt (ID: MD30)* "Advanced, planning, strategically-aware AI (as defined by Carlsmith) is developed"	*Status quo question (ID: STQ15)* By 2029, will an AI escape containment?
1	*Expert prompt (ID: HP27b)* "Large positive political or social changes e.g.: next gen of ML researchers have alignment-friendly culture"	*Status quo question (ID: STQ59)* In 2050, will the general consensus among experts be that the concern over AI risk in the 2020s was justified?

Conceptual uniqueness score	**0**	**0.2**	**0.5**	**0.8**	**1**
Death / harm	1	0	1	2	1
Power-seeking	0	0	0	5	0
Agenda progress	0	0	5	0	0
General alignment	0	1	2	4	1
Political response	0	0	4	0	0

Developer perception	0	0	0	1	3

Table A3.2.6: *Conceptual uniqueness scores by theme*

Operationalization uniqueness rating

Below are examples of each of the ratings used, along with the status quo question judged most similar.

0.2	*AICT question (ID: PV50)* Before 2050, an AI system is the cause of the death of at least 100 people OR is the cause of >$1 billion in economic damage within a 1-year period and subsequent analyses can not attribute the incident entirely to human error/malice.	*Status quo question (ID: 10)* Before 2032, will we see an event precipitated by AI malfunction that causes at least 100 deaths and/or at least $1B 2021 USD in economic damage?
0.5	*AICT question (ID: JL30a)* Resolves positively if 1 AND 2 resolve positively: 1) Suppose we commit, in 2030, to running a survey of the leadership of AI alignment research organizations which existed in or before 2022, asking: "By how much (%) did the research done by the following organizations and/or by independent alignment researchers between 2022 and 2030 decrease the chance that unaligned TAI is built before 2100?" This survey yields a median response of <=10%. “AI alignment research organizations” are listed below. Aligned AI; Anthropic; ARC; Future of Humanity Institute; MIRI; Redwood Research 2) By the end of 2030, no paper produced by one of the alignment organizations listed above has received an Outstanding Paper prize at NeurIPS, ICML, or ICLR.	*Status quo question (ID: 205)* Will OpenAI announce that it has solved the core technical challenges of superintelligence alignment by June 30, 2027?
0.8	*AICT question (ID: ZD40)* Resolves positively if (1) AND (2) are true. 1) By 2040, privately developed and deployed AI systems have caused accidental damage of at least (a) the equivalent of 0.001% of the world GDP for 2040, or (b) 1,000 human deaths, where “accidental” means “unintended by the groups that designed or deployed the AI systems”. 2) Between 2037 and 2040, and conditional on (1), the total amount of AI-caused damage over that period is within 30% of the highest 3-year period between 2030 and 2040, for whichever of financial damage or deaths satisfied the criteria in (1).	*Status quo question (ID: 10)* Before 2032, will we see an event precipitated by AI malfunction that causes at least 100 deaths and/or at least $1B 2021 USD in economic damage?
1	*AICT question (ID: MD30)* By 2030, Joe Carlsmith says he has ≥90% confidence that an AI fitting his APS definition exists. If Joe Carlsmith is	*Status quo question (ID: 15)* By 2029, will an AI escape containment?

	unavailable to offer resolution, we will ask a panel of experts nominated by 2022 Open Philanthropy researchers.	Market resolves YES if an AI living on a server ends up also being run on another server, and there is strong evidence that this was not a result of a human deliberately copying the AI, and the AI was not intended to do this.

Operationalization uniqueness score	**0**	**0.2**	**0.5**	**0.8**	**1**
Death / harm	0	1	2	2	0
Power-seeking*	0	0	1	1	2
Agenda progress	0	0	2	3	0
General alignment	0	0	3	5	0
Political response*	0	0	1	2	0
Developer perception	0	0	0	2	2

Table A3.2.7: *Operationalization uniqueness scores by theme.*
**These themes contained one expert prompt that was not operationalized due to time constraints.*

Conjunctive uniqueness

Some expert nodes were unique in the way that they combined different subject matter themes together into a conjunctive question. Most status quo questions are “simple” in this regard—while they may contain multiple different conditions, it is relatively unusual for them to contain conditions on more than one theme. Without information on the relationship between two or more conditions, it is impossible to determine whether their value together is additive, synergistic, or antagonistic. It is therefore very difficult to interpret independent information on each of the conditions as a “whole picture.”

Expert-chosen conjunctive questions do the job of highlighting which themes and conditions are likely to have a useful relationship to one another (much as conditional trees themselves do). Conjunctive questions also constitute a large part of possible question space, and failing to explore them could leave a large number of the highest value questions undiscovered. Indeed, it seems intuitively likely that most of the very highest value questions are conjunctive—a high value question can usually be made at least slightly more valuable by combining it with another question containing independent information.

Conditions in conjunctive questions can be joined with AND, OR, or XOR (though the last, exclusive OR, was never used by our expert sample or question-writing team). AND relations narrow the predictive target, relative to each condition independently, and are frequently used when combining multiple weak signals (high sensitivity, low specificity) for a stronger signal. OR relations broaden the target, and might be used when a strong signal (high specificity, low sensitivity) doesn’t on its own capture enough relevant scenarios.

A list of conjunctive questions from the AICT Alignment set with a unique combination of themes is below (a total of 10, out of 25 Alignment questions). The dominance of AND operators in this set suggests a preference for using conjunctions to narrow the prediction target.

	Theme 1	Theme 2	Conjunction operator
HP27a	General alignment	Agenda progress	OR
CX70	General alignment	AGI	AND
ZZ32	General alignment	AI companies	AND
VL70	General alignment	Political response; Competitive environment	AND
PV50	Death / harm	Agenda progress	AND
ZD30	Death / harm	Power-seeking	OR
MD40	Death / harm	Political response	AND
JL50	Death / harm	Political response	AND
XS50	Developers	Compute	AND
HP27b	Developers	Political response	AND

Table A3.2.8: *Questions in the AICT alignment set combining two themes*

Appendix 4: VOI technical explanation

VOI (Value of Information). We use Kullback-Leibler divergence to measure how much of a difference crux c would make, whether it resolves positively (green below) or negatively (yellow). Kullback-Leibler is commonly used in mathematical statistics to measure how different one probability distribution is from another. You can think of $P(U|c)$ and $P(U)$as very spiky distributions, so it serves here as well. The subject also has an idea of how likely crux c is to resolve positively ($P(c)$); hence, we can compute KL divergence in expectation. VOI answers the question, "How much will I gain by knowing the outcome of c?" ("gain" being measured in expected log score on the ultimate question U).

Case where c resolves positively — Case where c resolves negatively

$$VOI_{log}(P(U),\ P(U|c),\ P(c)) = KL((P(U|c),\ P(U)) * P(c)) + KL((P(U|\neg\, c),\ P(U)) * (1\ -\ P(c)),$$

where $KL(A, B)$ is defined as $A * log\left(\frac{A}{B}\right) + (1-A) * log\left(\frac{1-A}{1-B}\right)$

Case where A resolves positively (where U resolves positively given condition) — Case where A resolves negatively (where U resolves negatively given condition)

VOI is constrained by how small or large the subject's unconditional $P(U)$ is. If their $P(U)$ is very small, they're quite certain to begin with, and even the most valuable possible thing they could learn must necessarily be pretty low VOI. For example, if Adeeb's unconditional $P(U)$ is one-in-one-million, he believes that there is only a one-in-a-thousand chance he'll *ever* learn something that will update his $P(U)$ as high as 0.1%; a one-in-one-million chance he'll ever learn something that will update him to 100%. This latter is the hypothetical most valuable thing, the maximum possible VOI for Adeeb, i.e., the piece of information that would determine U. To level the playing field, we use percent-of-max VOI (POM VOI), which tells us the VOI of a crux, for a subject, relative to the maximum VOI[72] for that subject. If a subject is very certain to begin with (i.e. their $P(U)$ is very close to 0% or 100%), they have less to gain by learning the answer than another subject who is less certain.

VOD (value of discrimination) also employs KL divergence to examine the difference crux c is expected to make in the disagreement between two subjects a and b. They currently disagree by D_{init}; basically, this orange term is how much a would gain by believing b if b were right about $P(U)$, and vice versa, since we don't know who is right. Likewise, the green term is how much a would gain by believing b if b were right about $P(U|c)$ and vice versa, and the yellow term is the same for $P(U|\neg c)$. We weight these terms to get disagreement in expectation, using geometric mean of odds to arrive at a compromise between their two ideas about $P(c)$. Finally, VOD is the simple difference between the subjects' current

[72] Some readers may recognize maximum VOI given P(U) as the entropy of U. Likewise, KL divergence is also known as relative entropy. KL(P(U|c) || P(U)) is equivalent to the reduction H(U) - H(U|c) in entropy (denoted by H) conditional on the question resolving positively. VOI can therefore be interpreted as the fractional reduction in the entropy, or uncertainty, of U, provided by the answer to the crux question c.

disagreement and their expected disagreement. Negative VOD implies they expect to disagree more than they currently disagree; positive VOD implies they expect to disagree less.

$$D_{init} = \underbrace{KL(P_a(U),\ P_b(U)) + KL(P_b(U),\ P_a(U))}_{\text{Symmetric KL-divergence}}$$

$$\mathbb{E}[D] = (KL(P_a(U|c),\ P_b(U|c)) + KL(P_b(U|c),\ P_a(U|c))) * GMOD(P_a(c),\ P_b(c)) + (KL(P_a(U|\neg c),\ P_b(U|\neg c)) + KL(P_b(U|\neg c),\ P_a(U|\neg c)) * GMOD(P_a(\neg c),\ P_b(\neg c))$$

where $KL(A, B)$ is defined as $A * log\left(\frac{A}{B}\right) + (1 - A) * log\left(\frac{1 - A}{1 - B}\right)$

and $GMOD(A, B)$ is defined as $\frac{\sqrt{\frac{A}{1-A} \times \frac{B}{1-B}}}{\sqrt{\frac{A}{1-A} \times \frac{B}{1-B}} + 1}$

(i.e. convert probabilities to odds, take geometric mean, convert back to probabilities)

$$VOD_{log} = D_{init} - \mathbb{E}[D]$$

Similarly to VOI, VOD is constrained by how far apart two subjects begin. If we want to know how much of two people's disagreement a crux would resolve in expectation, we look to percent-of-max VOD (POM VOD), which we get simply by dividing VOD by their initial disagreement.

Appendix 5: Question combinations survey details

See above for an elicitation template.

The cohort of superforecasters who participated in our main question rating survey was asked to provide forecasts for a series of "scenario interactions," that is, a set of positive and/or negative resolutions to four forecasting questions. A scenario interaction was defined by a coding corresponding to the resolution of each individual question, denoted with either 1 or 0 for positive or negative resolution; this set would be considered as a whole for forecasting. For example, given questions A, B, C and D, a scenario interaction in which A and B resolved positively but C and D resolved negatively would be coded as A=1 B=1 C=0 D=0. Respondents were asked to estimate the probability of the scenario interaction as a whole, and the conditional probability of the ultimate question U given the scenario interaction. This allowed us to see whether the resolutions of certain questions correlated with each other, as well as how much independent information questions provided on P(U).

Respondents were given personalized documents containing answers they had given in previous elicitations for P(U), as well as for both the probability of individual questions and the conditional probability of U (in the template above these fields are marked with "XX"). The document also contained links to the operationalizations of each question. Below this, codes for scenario interactions were displayed next to yellow fields for the respondents to input forecasts for each scenario interaction. Respondents were shown scenario interactions in blocks that were shown in a random order. Below the scenario interaction elicitation we gave participants the opportunity to revise probabilities they had given on the questions individually.

Appendix 6: Process for Conducting Interviews

What we did in the AI conditional trees interviews

[5 mins] Intro

- Briefly explained what conditional trees were (with visual), what the rules of the exercise were, what the goals were
- Noted prize offered for best questions
- Noted confidentiality policy
- Checked if recording was ok

[5-10 mins] Preliminaries

- Elicited interviewee's initial P(U) (without probing reasoning at all)

- Checked on appropriate timescale for nodes
- Discussed what interviewee saw as main forces driving risk of U

[15-20 minutes] Node 1

- Brainstorming - Asked them to brainstorm events or scenarios which could occur by [node 1 timepoint], which are at least 10% likely, which would increase their P(U) a lot.
    - Sometimes if P(U) started very high, we would instead ask them to brainstorm negative updates.
    - If they're having trouble coming up with things, you can prompt them with "driving forces" they mentioned previously.
- Choosing a node - Read back the nodes they'd brainstormed to them, asked them to pick what they think is the best one (or best combination—they're allowed to have ANDs and ORs, as long as they end up with a yes/no question).
- Adding detail - Asked them to flesh the node out more, if they had any ideas for operationalization, or for concrete examples. Asked for P(U|c).
    - In retrospect, we also should have asked for P(c)!
- Used a mind-mapping tool to draw their conditional tree with P(U), node 1 and node 1's P(U|c).

[10-15 minutes] Node 2

- (As above)

[10-15 minutes] Node 3

- (As above)

[5 minutes] Wrap up questions

- These varied a lot, but included:
    - Asking them to rank other interviewees' nodes
    - How their answers would change if they were considering a different formulation of U
    - If they could name people or perspectives with whom they disagreed, but respected
    - Influences on their views

What we thought worked well about the interview process

- Having a bit of more open-ended discussion before diving into the nodes seemed to help interviewees get warmed up, and helped us as interviewers understand more of the context for their views.

- Showing participants a visual representation of their tree as they were creating it was a helpful aid (and visuals in general tend to be helpful).

- While a few participants wanted to talk longer than an hour, an hour seemed like a pretty good bar for the amount of time experts were happy to give us.

- Telescoping through different timepoints the way we did in the interview was novel for many experts, and several commented that the interview format prompted them to think differently.

- “People I disagree with but respect” was interesting, and could have been useful for finding further interview subjects (though I don’t remember if we in fact found any subjects this way).

What didn’t work well about the interview process

- We were asking them to do quite a lot in an hour, and frequently ran out of time for interview segments we’d planned.
- We initially started with a timescale that was miscalibrated to our expert samples’ beliefs, and as a result we got less useful material from some interviews.
- Sometimes interviewers struggled to get interviewees to conform to the format, and when this was the case it was much harder (sometimes impossible) to operationalize the material. (By and large these “non-compliant” interviewees were older professors.)
- There was a general tendency toward proposing dramatic low probability things rather than more likely things that were less dramatic, and our interview technique may have encouraged that to some extent.
- Asking interviewees to rank other people’s nodes was pretty time-intensive, so we dropped this.

Things we’re still uncertain about

- While we elicited nodes going “up” the tree, that is, starting at the earliest timepoint, we could have also gone “down” the tree and asked people for nodes starting at the latest timepoint. People would probably give pretty different answers this way, but it’s not clear whether this would be better or worse. It’s probably worth testing this out at some point to see if there are interesting effects.
- We specified a probability threshold for nodes as being >=10% likely, but there’s an argument that this threshold should be higher. Sometimes experts assign higher probability to risk events in their domain than superforecasters do, and if you’re targeting “events a superforecaster would think are 10% likely,” you may need to skew upwards in expert elicitation. (It’s not clear to me that we should be targeting a superforecaster bar for many conditional tree use cases.)

  There’s another argument along the lines that 10% likelihood is not a good threshold if interviewees anchor to it, because higher probability ranges may constitute richer search spaces for high VOI questions.

  In practice, we would sometimes slightly tweak questions at the operationalization stage so that they would be more probable. This usually also made them simpler, because we were e.g. removing conditions, and this reduces cognitive load on forecasters.
- Although our default was to ask people to move up a strict “increasing risk” path, if someone started with a very high estimate or got to a high estimate quickly with early nodes, we would ask them to switch to “decreasing risk.” We didn’t have super principled reasons for either starting with “increasing risk” or switching. But experimenting with the “sign” of elicitation could be interesting.

- You could decrease the average time per interview by 10-15 minutes by omitting the third node, and possibly this is worth it (though note that the marginal cost here is relatively small).

Miscellaneous recommendations

- Avoid getting too into the weeds on the mechanics of conditional trees—most of them aren't relevant to the interviewee's task.
- Ideally you should know what the modal timescale in the field is for U, so you can pick an appropriate and consistent set of timepoints for the exercise. If you don't have this information or if there's very little consensus in the field, it might be better to figure out an appropriate timescale on a person-by-person basis.
- Avoid asking leading questions (e.g., ask "How would you describe _?", not "Would you describe _ as [something the interviewee hasn't offered]?")
- Take copious notes in interview, and in the hour after the interview go back and add important details you didn't have time to record during. This is really important to do while your memory is fresh, and will help with operationalization later!
- Don't try to fill silences—leave them open for the interviewee to talk. Past the intro stage, the interviewee should do the vast majority of the talking, with you mostly offering gentle steering or asking clarifying questions.
- I strongly recommend preparing a note-taking doc ahead of time! Since you'll be repeating some of what the interviewee says back to them, you'll need to write these parts up in a way that's easy for you to find and read. The note-taking doc can also have prompts for you to remind you of target times or things you're saying.
- There are various mapping tools around that you could use for the conditional tree visual. I used draw.io. You'll have to share your screen in order to show interviewees the diagram.

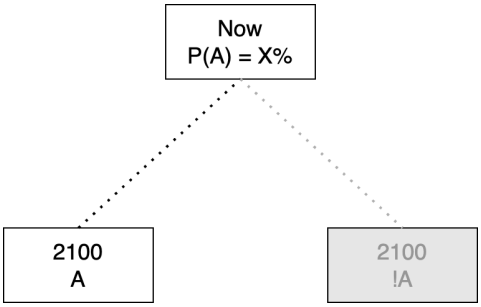
Example script

Intro

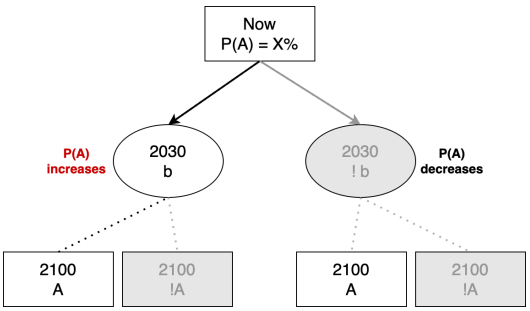
Time target: 5 minutes

Thanks for taking the time to meet with me. I'm going to start by briefly explaining what a conditional tree is, and the goals of this exercise. Please feel free to jump in with questions at any point.

A conditional tree is a web linking near-future events to far-future events probabilistically. To illustrate why such a tool might be useful... [Show fig 1]



Say you have a prediction about a far-future event, “A”, happening by 2100. Currently, you believe the probability of “A” is X%. So what’s on the path to “A” occurring or not occurring? What could happen before “A” that would cause you to update on its probability? [Show fig 2]



A conditional tree represents a sequence of these events, each providing further evidence on the future state of “A”. Each intermediary event is framed as a “yes/no” question whose answer changes the probability of “A” by some magnitude. These events also need to be somewhat likely to happen—at least 10% likely, in your judgment.

[Check if they have any questions. You can optionally provide a pre-prepared concrete example conditional tree here.][73]

Today we’re going to construct your conditional tree for [outcome].[74] We’re hoping that this exercise will help us find the best leading indicators for [outcome], which in turn will help forecasting concentrate on the most valuable questions in [field]. We think it will also be useful for mapping out the views of different experts on this topic, and seeing where there is consensus and disagreement.

[Note here if you’re offering any incentive for performance. Avoid getting into the weeds unless they ask.]

Your identity and the content of this interview will be confidential, and this information will only be shared with other members of the research team where necessary. With that in mind, would you be comfortable with me recording this interview?

Okay, let’s jump in. *[Start recording if permission given.]*

---

[73] Here I’d definitely avoid subject matter similar to what you’ll be asking about in the interview, to avoid priming them.

[74] Not everyone will ask, but it’s good to have a solid (non-verbose) operationalization prepared for U in case they do. You can also emphasize that because forecasting U isn’t a goal of the exercise, they can consider U in a common sense way.

Preliminaries

Time target: 5 minutes

Currently, what do you think is the probability of [outcome] by [outcome date]?[75] *[Record this on the conditional tree visual.]*

What do you see as the main driving forces that will increase or decrease this probability over the next few decades?[76]

[If you have a flexible tree timeline, ask what an appropriate one would be for your interviewee, perhaps providing a few different options.]

Node

Time target for each node: 15 minutes[77]

Now we're going to construct your conditional tree, starting with [timepoint 1]. First we'll do some brainstorming. What event or scenario could occur by [timepoint 1] that would most [increase/decrease] your probability of [outcome]? Remember, it must be somewhat likely to happen, at least 10% in your judgment.

[After 5 to 10 minutes of brainstorming, read back the options they listed.] Out of these, which do you think is the best? If you want, you can combine more than one.[78]

[Once they've chosen, ask:] What do you think is the probability of [node 1] by [timepoint 1]?

Remember that your current probability of [outcome] is [P(U)]. Supposing [node 1] happened by [timepoint 1], what would be your updated probability of [outcome] by [end date]?

[Record node 1 and P(U|c) on the conditional tree visual and show this to them.]

[If there's time remaining in this timeslot, ask for more detail on the node or ideas for operationalization.]

[Repeat for each node.]

Wrap-up

Time target: 5 minutes

---

[75] Sometimes people are reluctant to give numbers. Here it can be helpful to say that you're looking for a "rough" or "provisional" number, and that no one is going to hold them to it, it's just useful as a baseline for this exercise. But do push for an actual point estimate–it makes the rest of the exercise easier. At this stage you want something that's unconditional and all-things-considered, so don't accept "probability of [U] if [such-and-such] is true" as a baseline P(U). (A common form of this is, "If we do nothing from this point on, then the bad outcome is very probable!")

[76] Take good notes and ask questions here to make sure you understand. This step is useful for providing the basic outlines of their models, so that when they get into the nitty gritty of node brainstorming you can follow their train of thought better.

[77] This assumes a 3-node tree. A 2-node tree could spend ~20 minutes per node.

[78] If they find choosing difficult, it might be helpful to clarify what "best" means in this context by reminding them that it's the highest information signal. Alternatively, you can try quickly eliciting P(c) and P(U|c) for several, and either input it into a VOI calculator or just go off a heuristic sense, if you feel comfortable doing that.

Thanks again for participating in the exercise—this is a really helpful contribution to our project, and I've enjoyed hearing your thoughts on [topic]. To wrap up, I just have a few quick questions.

Firstly, who would you say has most influenced your views on this topic?

Can you name any people or perspectives with which you disagree on this topic, but which you respect?

[If there's going to be any followup, mention that here. Otherwise finish by saying that the FRI team will operationalize their nodes, and then thank them again (or whatever polite and respectful signoff you prefer).]

Appendix 7: Make forecasts on questions from this report

At the following links, you can make your own predictions on the ultimate question and a sample of the most informative conditional questions from the AICT question set.

Manifold

- **Ultimate question**: Before 2100, will AI cause the human population to fall below 5000 individuals?
- **CX30**: Before 2030, will there be an AI-caused administrative disempowerment?
- **VL30**: Between 2023 and 2030, will revenue from deep learning double every two years?
- **ZA50**: Before 2050, will an AI system be shut down due to exhibiting power-seeking behavior?
- **CX50**: If, before 2050, AI kills more than 1 million people, will the policy response be insufficient?
- **Conditional pairs**: Conditional on the following events, before 2100, will AI cause the human population to fall below 5000 individuals?

Metaculus

The questions listed above can also be forecast on Metaculus. Find them all on this question page.