Forecasting Research Institute

Roots of Disagreement on AI Risk:

Exploring the Potential and Pitfalls of Adversarial Collaboration

Authors: Josh Rosenberg, Ezra Karger, Avital Morris, Molly Hickman, Rose Hadshar, Zachary Jacobs, Philip Tetlock

Roots of Disagreement on AI Risk: Exploring the Potential and Pitfalls of Adversarial Collaboration[1]

Josh Rosenberg,[*] Ezra Karger,[‡,*] Avital Morris,[*] Molly Hickman,[*] Rose Hadshar,[*] Zachary Jacobs,[*] Philip Tetlock[‡,*]

Abstract

We brought together generalist forecasters and domain experts (n=22) who disagreed about the risk AI poses to humanity in the next century. The "concerned" participants (all of whom were domain experts) predicted a 20% chance of an AI-caused existential catastrophe by 2100, while the "skeptical" group (mainly "superforecasters") predicted a 0.12% chance. Participants worked together to find the strongest near-term cruxes: forecasting questions resolving by 2030 that would lead to the largest change in their beliefs (in expectation) about the risk of existential catastrophe by 2100. Neither the concerned nor the skeptics substantially updated toward the other's views during our study, though one of the top short-term cruxes we identified is expected to close the gap in beliefs about AI existential catastrophe by about 5%: approximately 1 percentage point out of the roughly 20 percentage point gap in existential catastrophe forecasts. We find greater agreement about a broader set of risks from AI over the next thousand years: the two groups gave median forecasts of 30% (skeptics) and 40% (concerned) that AI will have severe negative effects on humanity by causing major declines in population, very low self-reported well-being, or extinction.

[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 for data analysis and editing on the report, Taylor Smith and Bridget Williams as adversarial collaboration moderators, and Kayla Gamin, Coralie Consigny, and Harrison Durland for their careful editing. We thank Elie Hassenfeld, Eli Lifland, Nick Beckstead, Bob Sawyer, Kjirste Morrell, Adam Jarvis, Dan Mayland, Jeremiah Stanghini, Jonathan Hosgood, Dwight Smith, Ted Sanders, Scott Eastman, John Croxton, Raimondas Lencevicius, Alexandru Marcoci, Kevin Dorst, Jaime Sevilla, Rose Hadshar, Holden Karnofsky, Benjamin Tereick, Isabel Juniewicz, Walter Frick, Alex Lawsen, Matt Clancy, Tegan McCaslin, and Lyle Ungar for comments on the report.

* Forecasting Research Institute

‡ Federal Reserve Bank of Chicago

‡ Wharton School of the University of Pennsylvania

Executive summary

In the summer of 2022, researchers affiliated with the Forecasting Research Institute (FRI) (a)[2] ran the Existential Risk Persuasion Tournament (XPT) (a), which identified large disagreements between domain experts and generalist forecasters about key risks to humanity (Karger et al. 2023). This new project—a structured adversarial collaboration run in April and May 2023—is a follow-up to the XPT focused on better understanding the drivers of disagreement about AI risk.

Methods

We recruited participants to join "AI skeptic" (n=11) and "AI concerned" (n=11) groups that disagree strongly about the probability that AI will cause an existential catastrophe by 2100.[3] The skeptic group included nine superforecasters and two domain experts. The concerned group consisted of domain experts referred to us by staff members at Open Philanthropy (the funder of this project) and the broader Effective Altruism community.

Participants spent 8 weeks (skeptic median: 80 hours of work on the project; concerned median: 31 hours) reading background materials, developing forecasts, and engaging in online discussion and video calls. We asked participants to work toward a better understanding of their sources of agreement and disagreement, and to propose and investigate "cruxes": short-term indicators, usually resolving by 2030, that would cause the largest updates in expectation to each group's view on the probability of existential catastrophe due to AI by 2100.

Results: What drives (and doesn't drive) disagreement over AI risk

At the beginning of the project, the median "skeptic" forecasted a 0.10% chance of existential catastrophe due to AI by 2100, and the median "concerned" participant forecasted a 25% chance. By the end, these numbers were 0.12% and 20% respectively, though many participants did not attribute their updates to arguments made during the project.[4]

We organize our findings as responses to four hypotheses about what drives disagreement:

---

[2] To ensure the stability of links in this report, we include stable archive.org links in parentheses after each citation to an external URL.

[3] We defined an "existential catastrophe" as an event where one of the following occurs: (1) Humanity goes extinct; or (2) Humanity experiences "unrecoverable collapse," which means either: (a) a global GDP of less than $1 trillion annually in 2022 dollars for at least a million years (continuously), beginning before 2100; or (b) a human population remaining below 1 million for at least a million years (continuously), beginning before 2100.

[4] For example, three out of six "concerned" participants who updated downward during the project attributed their shift to increased attention to AI risk among policymakers and the public after the release of GPT-4. For more details on the reasons for all updates, see the "Central Disagreement" section below and Appendix 4.

Hypothesis #1 - Disagreements about AI risk persist due to lack of engagement among participants, low quality of participants, or because the skeptic and concerned groups did not understand each others' arguments[5]

We found moderate evidence against these possibilities. Participants engaged for 25-100 hours each (skeptic median: 80 hours; concerned median: 31 hours), this project included a selective group of superforecasters and domain experts, and the groups were able to summarize each others' arguments well during the project and in follow-up surveys. (More)

Hypothesis #2 - Disagreements about AI risk are explained by different short-term expectations (e.g. about AI capabilities, AI policy, or other factors that could be observed by 2030)

Most of the disagreement about AI risk by 2100 is not explained by indicators resolving by 2030 that we examined in this project. According to our metrics of crux quality, one of the top cruxes we identified is expected to close the gap in beliefs about AI existential catastrophe by about 5% (approximately 1.2 percentage points out of the 22.7 percentage point gap in forecasts for the median pair) when it resolves in 2030.[6] For at least half of participants in each group, there was a question that was at least 5-10% as informative as being told by an oracle whether AI in fact caused an existential catastrophe or not.[7] It is difficult to contextualize the size of these effects because this is the first project applying question metrics to AI forecasting questions that we are aware of.

However, near-term cruxes shed light on what the groups believe, where they disagree, and why:

- **Evaluations of dangerous AI capabilities are relevant to both groups.** One of the strongest cruxes that will resolve by 2030 is about whether METR (formerly known as ARC Evals) (a) or a similar group will find that AI has developed dangerous capabilities such as autonomously replicating and avoiding shutdown. This crux

[5] Scott Alexander, among other XPT readers, suggested this possibility: "Many of the people in this tournament hadn't really encountered arguments about AI extinction before (potentially including the "AI experts" if they were just eg people who make robot arms or something), and a couple of months of back and forth discussion in the middle of a dozen other questions probably isn't enough for even a smart person to wrap their brain around the topic". See Scott Alexander, "The Extinction Tournament", *Astral Codex Ten, (*July 20, 2023*)* https://www.astralcodexten.com/p/the-extinction-tournament (a).

[6] The best convergent crux, "ARC Evals," would narrow the disagreement between the median pair from 22.7 percentage points to 21.48 percentage points in expectation, which means eliminating 5.35% of their disagreement. Note that this statistic refers to the median pair by POM VOD. See "ARC Evals" for more details. For magnitudes of value of information effects, see here.

[7] For more details, see "Contextualizing the magnitude of value of information". In more concrete terms, this is equivalent to a forecasting question with the following characteristics:

- A concerned participant with original P(AI existential catastrophe (XC) by 2100) = 25% identifies a crux that has: P(crux) = 20%, P(AI XC|crux) = 6.2%, and P(AI XC|¬crux) = 29.7%
- A skeptic participant with original P(AI XC by 2100) = 1% identifies a crux that has: P(crux) = 20%, P(AI XC|crux) = 3.37%, and P(AI XC|¬crux) = 0.41%

illustrates a theme in the disagreement: the skeptic group typically did not find theoretical arguments for AI risk persuasive but would update their views based on real-world demonstrations of dangerous AI capabilities that verify existing theoretical arguments. If this question resolves negatively then the concerned group would be less worried, because it would mean that we have had years of progress from today's models without this plausible set of dangerous capabilities becoming apparent. (More)

- **Generally, the questions that would be most informative to each of the two groups are fairly distinct.** The concerned group's highest-ranked cruxes tended to relate to AI alignment and alignment research. The skeptic group's highest-ranked cruxes tended to relate to the development of lethal technologies and demonstrations of harmful AI power-seeking behavior. This suggests that many of the two groups' largest sources of uncertainty are different, and in many cases further investigation of one group's uncertainties would not persuade the other. (More)
- **Commonly-discussed topics—such as near-term economic effects of AI and progress in many AI capabilities—did not seem like strong cruxes.** (More)

Hypothesis #3 - Disagreements about AI risk are explained by different long-term expectations

We found substantial evidence that disagreements about AI risk decreased between the groups when considering longer time horizons (the next thousand years) and a broader set of severe negative outcomes from AI beyond extinction or civilizational collapse, such as large decreases in well-being or total population.

Some of the key drivers of disagreement about AI risk are that the groups have different expectations about: (1) how long it will take until AIs have capabilities far beyond those of humans in all relevant domains; (2) how common it will be for AI systems to develop goals that might lead to human extinction; (3) whether killing all living humans would remain difficult for an advanced AI; and (4) how adequately they expect society to respond to dangers from advanced AI.[8]

Supportive evidence for these claims includes:

- Both groups strongly expected that powerful AI (defined as "AI that exceeds the cognitive performance of humans in >95% of economically relevant domains") would be developed by 2100 (skeptic median: 90%; concerned median: 88%). Though, some skeptics argue that (i) strong physical capabilities (in addition to cognitive ones) would be important for causing severe negative effects in the world, and (ii) even if AI can do most cognitive tasks, there will likely be a "long tail" of tasks that require humans.
- The two groups also put similar total probabilities on at least one of a cluster of bad outcomes from AI happening over the next 1000 years (median 40% and 30% for

[8] See "Understanding each other's arguments" and Appendix 10 for additional discussion of key areas of disagreement.

concerned and skeptic groups respectively).[9] But they distribute their probabilities differently over time: the concerned group concentrates their probability mass before 2100, and the skeptics spread their probability mass more evenly over the next 1,000 years.

- We asked participants when AI will displace humans as the primary force that determines what happens in the future.[10] The concerned group's median date is 2045 and the skeptic group's median date is 2450—405 years later.

Overall, many skeptics regarded their forecasts on AI existential risk as worryingly high, although low in absolute terms relative to the concerned group.[11]

Despite their large disagreements about AI outcomes over the long term, many participants in each group expressed a sense of humility about long-term forecasting and emphasized that they are not claiming to have confident predictions of distant events.

Hypothesis #4 - These groups have fundamental worldview disagreements that go beyond the discussion about AI

Disagreements about AI risk in this project often connected to more fundamental worldview differences between the groups. For example, the skeptics were somewhat anchored on the assumption that the world usually changes slowly, making the rapid extinction of humanity unlikely. The concerned group worked from a different starting point: namely, that the arrival of a higher-intelligence species, such as humans, has often led to the extinction of lower-intelligence species, such as large mammals on most continents. In this view, humanity's prospects are grim as soon as AI is much more capable than we are. The concerned group also was more willing to place weight on theoretical arguments with multiple steps of logic, while the skeptics tended to doubt the usefulness of such arguments for forecasting the future.

---

[9] These outcomes were: AI causing extinction intentionally, unintentionally, or via misuse, misalignment causing a 50% drop in human population, or human well-being dropping to <4/10 because of AI misalignment, accidents, or misuse. These were phrased to be mutually exclusive outcomes. See "Survey on long-term AI outcomes" section and Appendix 5 for more details.

[10] The full question we asked was, "By what year, if ever, do you expect to agree with the following statement? 'AI has displaced humans as the primary force that determines what happens in the future. It now has at least as much power relative to humans as humans had relative to other species in 2023.'" Note that this would not necessarily be seen as a negative outcome by all participants.

[11] Note: All participant quotes have been regularized to American English to preserve anonymization. Participants classified as AI skeptics stated, for example, "Also, none of this is to say from a skeptic point of view the issues are not important[.] I think for us a 1% risk is a high risk;" "[T]he 'risk-concerned' camp (I'm using scare quotes because I consider that I'm risk concerned, even though technically I'm in the risk-skeptic camp because I assign a far lower probability to extinction by 2100 relative to some);" "AIs could (and likely will) eventually have massive power;" "That said, still perceive overall risk as "low at a glance but far too high considering the stakes["];" "To my mind, there should be no difference in the policy response to a 1% chance of 60% of humanity dying and a 25% chance—both forecasts easily cross the threshold of being 'too damn high'."

Results: Forecasting methodology

This project establishes clear quantifiable metrics for evaluating the quality of AI forecasting questions. And we view this project as an ongoing one. So, we invite readers to try to generate cruxes that outperform the top cruxes from our project thus far—an exercise that underscores the value of establishing comparative benchmarks for new forecasting questions. See the "Value of Information" (VOI) and "Value of Discrimination" (VOD) calculators (a) to inform intuitions about how these question metrics work. And please reach out to the authors with suggestions for high-quality cruxes.

Broader scientific implications

This project has implications for how much we should expect rational debate to shift people's views on AI risk. Thoughtful groups of people engaged each other for a long time but converged very little. This raises questions about the belief formation process and how much is driven by explicit rational arguments vs. difficult-to-articulate worldviews vs. other, potentially non-epistemic factors (see research literature on motivated cognition, such as Gilovich et al. 2002; Kunda, 1990; Mercier and Sperber, 2011).

One notable finding is that a highly informative crux for each group was whether their peers would update on AI risk over time. This highlights how social and epistemic groups can be important predictors of beliefs about AI risk.[12]

[12] This could be due to normative influence (because people defer to their social or intellectual peers), or, more likely in our view, informational influence (because they think that, if people whose reasoning they trust have changed their mind by 2030, it must be that surprising new information has come to light that informs their new opinion). Disentangling these pathways is a goal for future work.

Table of Contents

Glossary

ARC Evals
An organization, now called **METR** (Model Evaluation & Threat Research), that works on assessing whether cutting-edge AI systems could pose catastrophic risks to civilization. See "ARC Evals" for discussion of forecasts conditional on METR finding evidence of AI having the ability to autonomously replicate, acquire resources, and avoid shutdown before 2030.

Convergent crux
A question such that, conditional on it resolving, two people or groups will, in expectation, disagree less than they do now. See "Convergent Cruxes" for discussion of convergent cruxes found in this study.

Cross-camp pair
A pair consisting of one member of the skeptic group and one member of the concerned group. See "VOD" for discussion of questions that would narrow or widen disagreement for the median cross-camp pair when ranked by **VOD**, and "Differences of Opinion Within Groups" for discussion of each cross-camp pair's differences on one question.

Divergent crux
A question such that, conditional on it resolving, two people or groups will, in expectation, disagree more than they do now. See "Divergent Cruxes" for discussion of divergent cruxes found in this study.

Existential catastrophe
Defined in this study as an event in which at least one of the following occurs:

1. Humanity goes extinct
2. Humanity experiences "unrecoverable collapse," which means either:
    a. <$1 trillion global GDP annually [in 2022 dollars] for at least a million years (continuously), beginning before 2100; or
    b. Human population remains below 1 million for at least a million years (continuously), beginning before 2100.

Flash forecast
A forecast on which participants were recommended to spend approximately 10 minutes.

IC
Instrumental convergence, the hypothesized tendency for intelligent agents to develop similar sub-goals that are helpful for achieving most other goals, even if their ultimate goals are very different. In particular, sub-goals like acquiring resources, avoiding being killed/destroyed, and avoiding interference from other agents could be helpful for achieving a wide variety of other goals. See "ARC Evals" for discussion of forecasts conditional on a model having capabilities that might suggest instrumental convergence.

METR
Model Evaluation & Threat Research. See **ARC Evals**.

U
The "Ultimate question." In this study: "Will AI cause an **existential catastrophe** by 2100?"

VOD
Value of Discrimination (VOD) is a measure of how much knowing the answer to a question would change relative beliefs between individuals, in expectation. It is useful for measuring convergence and divergence in expected beliefs between individuals. See "VOD" for discussion of questions that would narrow or widen disagreement between the skeptic and concerned groups in expectation and Appendix 2 for an explanation of how VOD is calculated.

VOI
Value of Information (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. See "VOI" for discussion of informative questions and Appendix 2 for an explanation of how VOI is calculated.

P(U)
The probability that U, the ultimate question, occurs. In this case, the probability that AI causes an existential catastrophe by 2100.

P(C)
The probability that a potential crux question occurs. See Appendix 1 for a list of candidate cruxes, and "VOI: Results Tables and Figures" for the median participant in each group's P(C) for each crux.

POM
Percent of max. When we present VOI and VOD for each question, we also present how much of the maximum VOI or VOD it captured in order to contextualize the magnitude of the results. See "VOI" for discussion of POM VOI and "VOD" for discussion of POM VOD. See "ARC Evals" for an example of calculating POM VOD.

Background & Motivation

From June through October 2022, researchers affiliated with the Forecasting Research Institute (FRI) conducted the Existential Risk Persuasion Tournament (XPT) (a). A clear pattern emerged in its findings: AI domain experts thought extinction due to AI in the 21st century was much more likely than skilled generalist forecasters ("superforecasters") thought, and neither group persuaded the other much, despite working collaboratively and being incentivized to share persuasive arguments.[13] In addition, experts and superforecasters often agreed about short-term AI developments, while still disagreeing about the likelihood of extinction due to AI.[14]

In April and May of 2023, FRI ran a follow-up AI adversarial collaboration[15] project that aimed to figure out what drives disagreement about long-run AI risk. We aimed to get more time from a select group of high-quality participants and supported them with moderators, adversarial collaboration video calls, and seminar discussions with AI experts, among other activities. To support deep engagement, we kept this study small: eleven "skeptics" and eleven "concerned" participants.

We also designed the project to identify short-run indicators ("cruxes") resolving by 2030 that could help to diagnose reasons for disagreement and act as signals for the level of long-run AI risk.[16] While the XPT questions were chosen by our research team, this project asked the participants to collaborate to find the strongest cruxes, or short-run AI questions that would change beliefs about long-run AI risk the most in expectation.

---

[13] The median AI expert predicted a 12% chance of catastrophe and a 3% chance of human extinction due to AI by 2100. The median superforecaster predicted a 2.13% chance of catastrophe and a 0.38% chance of extinction due to AI. While experts predicted higher chances of all potential extinction risks than superforecasters did (including nuclear weapons and biorisks), the effect was much more pronounced in the case of AI. For more on lack of convergence, see Ezra Karger, et al., "Forecasting Existential Risks Evidence from a Long-Run Forecasting Tournament", *Forecasting Research Institute*, August 8, 2023, https://static1.squarespace.com/static/635693acf15a3e2a14a56a4a/t/64f0a7838ccbf43b6b5ee40c/1693493128111/XPT.pdf (a).

[14] For example, superforecasters predicted that an AI would first win an International Math Olympiad gold medal in 2035 while experts predicted 2030. See Karger et al., "XPT report" (a), page 156. For full relevant analysis, see "Relationship between short-run forecasting questions and longer-term disagreements" section on page 41.

[15] "Adversarial collaboration" protocols, often enforced by "neutral" umpires, encourage each side to demonstrate their capacity to fairly characterize, not caricature, the views of the other—and then to reach ex ante agreements on the types of data, observational or experimental, that would induce each side to move toward the other's position. For examples of adversarial collaborations and additional information, see "About", Penn Arts and Sciences Adversarial Collaboration Project, Accessed on February 9, 2024, https://web.sas.upenn.edu/adcollabproject/about/ (a).

[16] Note that, in some conversations about cruxes for AI risk, the word "crux" is used for questions that would lead to large updates even if highly unlikely (what we call "red flags"). In this project, we are focused on expected updates: we looked for cruxes that would be the most important in expectation, weighting how much difference they would make if they happened by how likely they are to happen.

So, why do thoughtful people disagree so strongly about AI risk? We organize our findings into four hypotheses about drivers of disagreement.

First, people who disagree may have not spent enough time engaging with each other or may not understand each other's arguments. Some of our readers suspected that superforecasters had not digested the main arguments for AI risk and would have been more concerned if they had, whereas others suspected that experts simply spent too much time talking to people who share their worldview and hadn't spent enough time talking to thoughtful skeptics.[17] If this hypothesis were true, we would expect that the two groups would agree more if there were enough high-quality engagement between them to understand each other's arguments.

Second, the disagreeing groups could have different predictions about short-term (by 2030) AI developments, such as how likely AI is to develop dangerous capabilities or which AI policies society is likely to adopt. If this hypothesis were true, we would expect the two groups to agree if we condition on specific AI-related developments. For example, if they disagreed about how long it will take until AI can write code to improve itself, but agreed that this development would mean serious danger for humanity, then we would expect them to agree on AI risk if we condition on AI improving itself. In this project, we asked participants to make many such conditional forecasts.[18]

Third, they could disagree about how AIs will develop or how society will respond in the longer term (through 2100 or beyond). Perhaps the groups cannot identify short-term AI outcomes that distinguish their risk models, but expect very different long-term AI trajectories.

Finally, they could have more fundamental worldview disagreements that go beyond AI. If they agree about most AI-related developments but continue to disagree about AI risk, there could be something else underlying their difference of opinion. There could be disagreements about how much they trust different categories of evidence or argumentation, or what they believe about human ingenuity and resilience, or any number of other topics that go beyond AI.

---

[17] For example, Scott Alexander stated that, "Many of the people in this tournament hadn't really encountered arguments about AI extinction before (potentially including the "AI experts" if they were just eg people who make robot arms or something), and a couple of months of back and forth discussion in the middle of a dozen other questions probably isn't enough for even a smart person to wrap their brain around the topic". See Alexander, "The Extinction Tournament" (a). Similarly, one XPT participant wrote, "I've been spending enough time on LessWrong that I mostly forgot the existence of smart people who thought recent AI advances were mostly hype. I was unprepared to explain why I thought AI was underhyped in 2022". See Peter McCluskey, "Existential Risk Persuasion Tournament", *Less Wrong* (July 17, 2023) https://www.lesswrong.com/posts/YTPtjExcwpii6NikG/existential-risk-persuasion-tournament#Persistent_Disagreement_about_AGI (a).

[18] When eliciting conditional forecasts, the prompt given to participants read: "Conditional on this question resolving positively (by 2030), what is your probability that AI causes an existential catastrophe by 2100?"

How did we test potential drivers of disagreement?

We brought together 22 participants who disagreed strongly on AI existential risk. Half of the participants were termed AI "skeptics,"[19] people whose XPT forecasts of the probability that AI would cause extinction by 2100 were <1%, and who produced high-quality rationales for their forecasts. This group of 11 AI skeptics included nine superforecasters and two domain experts. The other 11 participants were people concerned about AI, whom we expected to forecast a >10% chance that AI would cause an existential catastrophe by 2100. The "AI concerned" participants were AI safety researchers and AI-knowledgeable generalist researchers who were recommended as being able to present and discuss AI-concerned views clearly.

We asked these two groups to engage deeply with each others' arguments and to work together to identify cruxes with the most potential to update their forecasts on AI existential risk.

Participants made an initial forecast on the core question they disagreed about (we'll call this U, for "ultimate question"): by 2100, will AI cause an existential catastrophe? We defined "existential catastrophe" as an event in which at least one of the following occurs:

1. Humanity goes extinct
2. Humanity experiences "unrecoverable collapse," which means either:
    c. <$1 trillion global GDP annually [in 2022 dollars] for at least a million years (continuously), beginning before 2100; or
    d. Human population remains below 1 million for at least a million years (continuously), beginning before 2100.

For additional resolution details, such as the definition of "cause," see Appendix 1.

Over the next eight weeks, participants made forecasts on candidate crux questions that could help explain the disagreement, generated new possible cruxes during adversarial collaboration calls, and debated and refined their reasoning on an online platform. (See section below for more details on how the project worked.)

The central disagreement

The two groups were selected for disagreeing strongly about the likelihood of existential catastrophe due to AI by 2100, and they continued to disagree throughout the project. At the outset, the median skeptic forecasted a 0.10% chance of existential catastrophe due to AI by 2100, and the median concerned participant forecasted a 25% chance. Over the course of

[19] Note: many people in the "skeptic" group describe themselves as concerned about risks from advanced AI, including but not limited to the risk of extinction, despite thinking those risks are less likely to materialize than the "concerned" group expects. For example, "Also, none of this is to say from a skeptic point of view the issues are not important[.] I think for us a 1% risk is a high risk." (Gus); "... the 'risk-concerned' camp (I'm using scare quotes because I consider that I'm risk concerned, even though technically I'm in the risk-skeptic camp because I assign a far lower probability to extinction by 2100 relative to some)" (Blake).

the two-month project, there was mild convergence: the skeptic group's median moved from 0.10% to 0.12% and the concerned group's median fell from 25% to 20%.

However, April–May 2023 was an exciting time in real-world AI developments: GPT-4 had just become available, and regulators and the public were beginning to respond. Several participants attributed their updated probability of extinction due to AI by 2100 to these developments, and not to updates they made based on their work on this project.[20]

	Mean	**Median**	**Range**
Skeptic	0.54%	0.1%	0.0000001% - 3%
Concerned	28.4%	25%	4% - 65%

Table 1. Group P(AI-caused existential catastrophe by 2100), based on each participant's initial forecast

	Mean	**Median**	**Range**
Skeptic	0.46%	0.12%	0.0001% - 2%
Concerned	23.8%	20%	2.4% - 55%

Table 2. Group P(AI-caused existential catastrophe by 2100) at the end of the project

Six people in the concerned group lowered their forecasts and none raised them. Five people in the skeptic group raised their forecast, four lowered, and one raised but only because of an initial typo. For details on updated forecasts and reasons for updates from each participant, see Appendix 4.

[20] For full details, see Appendix 4. Six out of the 11 concerned participants updated downward during the project. Three out of those six cited policy responses as the reason for their updates, one cited an improved understanding of the base rate of non-human extinction after humans arose, one shifted some probability mass toward AI "takeover" rather than AI-caused existential catastrophe, and one did not explain their reasons for updating. Example quotes from participants citing policy responses as the reason for updating:

- "I have updated my prognosis to 30% [down from 60%], partially driven by positive updates in the area of point 4 making coordination and slowdown/stop of capability research more likely. This largely refers to the shift in public consciousness and the [O]verton window around the topic as I have perceived it over the past months, currently culminating in a public statement by most of the leading figures."
- "Slightly lowering my forecast [from 25% to 20%] as [relevant people take the risk seriously] has exceeded my (fairly high) expectations over the last couple of months."
- "I think my main update here [moving from 21% to 18%] has come from thinking a bit more deeply about AI regulation and what measures society will adopt to prevent catastrophes. I did not really include this as part of my original model, but it now seems somewhat likely that at least the EU and US will adopt some regulation that meaningfully reduces risk."

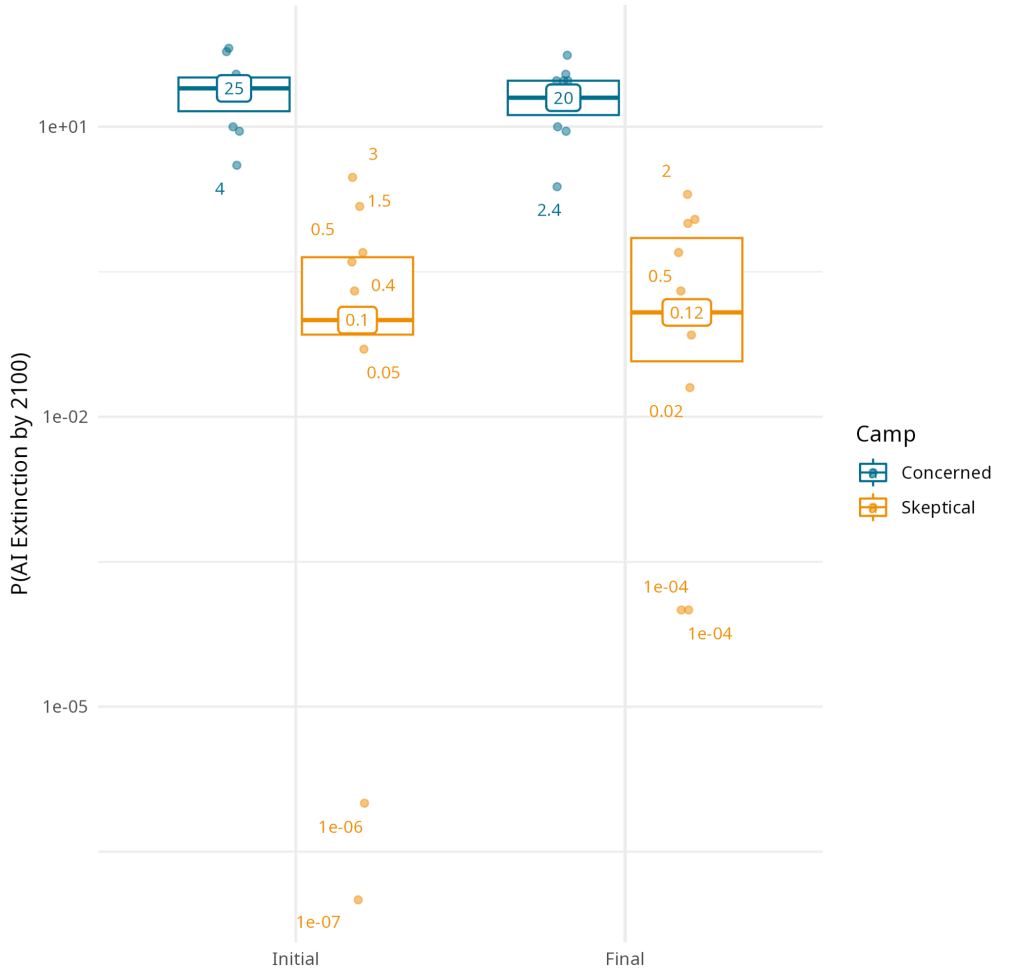

Figure 1. Initial and final P(AI existential catastrophe by 2100) for skeptic and concerned groups. See Appendix 4 for reasons for updates.

When we asked participants for forecasts on AI causing the deaths of more than 60% of the human population (within a 5-year period) before 2100, forecasts were closer than those about existential catastrophe, but there was still a large disagreement. The median concerned participant forecasted 32%, and the median skeptic forecasted 1%. This supports the claim that skeptics think that AI killing many people is more likely than causing existential catastrophe, whether because of the likelihood that it is useful to some other goal, the difficulty of killing people living in remote areas, or the likelihood of successful societal response to extreme catastrophe. However, the disagreement between groups is still large. Their disagreement about AI risk is deeper than the question of whether a very small number of humans will survive an AI catastrophe.

Although the disagreement between groups was large, many participants emphasized that their forecasts should be taken with a sense of humility, and that long-run forecasting is inherently uncertain and they are not claiming to have complete pictures of how events will

unfold over the coming decades.[21] Most previous evidence on judgmental forecasting applies to geopolitical forecasts on 0-2 year time horizons.[22]

[21] For example, one participant described their forecast as based on a "*very* rough back-of-the-envelope estimate" (Stella) and another said, "I'm with Tetlocks original view that long-term forecasts of this nature are very unreliable" (Gus). Skeptics who were not subject-matter experts were particularly candid when they were forecasting questions that involved technical details. On a question about the lowest price of GFLOPs, one skeptic said "I'm operating completely outside of my area of expertise here, so no one should hesitate to correct me" (Blake), and another said "This is very far away from my area of understanding. Mostly running on crude estimates of current trends with some leeway in the nearer term for newer hardware designed specifically optimized for reducing the cost of AI training" (Eve).

[22] 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 (a). 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, (a).

How the AI adversarial collaboration worked

The core activities of this project ran from April 1 to May 31, 2023.

Recruitment

We recruited 11 participants for the skeptic group who forecasted <1% on P(AI extinction by 2100) in the XPT, and stood out to either our research team or other XPT participants as having high-quality rationales and being collaborative. Nine out of 11 of these participants were superforecasters and two were domain experts from our XPT sample.[23]

We recruited 11 participants for the concerned group whom we expected to forecast >10% on P(AI existential catastrophe by 2100) and to be collaborative communicators. We began with recommendations for participants from staff members at Open Philanthropy and then did a broader search for reputable AI safety researchers and AI-knowledgeable generalist researchers (such as participants from Rethink Priorities (a) and Epoch (a)). Several of the concerned participants also had strong public track records of forecasting accuracy on short-run questions.

In the rest of this report, to preserve anonymity, we refer to participants with assigned aliases. Aliases beginning with A-K are assigned to skeptics, and aliases beginning with P-Z are assigned to concerned participants (in random order within each group).

Activities to facilitate engagement between the skeptic and concerned groups

As preparation for the project, we asked the skeptic group to read Holden Karnofsky's Most Important Century series (a) and related resources on AI existential risk recommended by staff members at Open Philanthropy.

Most discussion between groups happened on an online forum and forecasting platform that we set up for this project. Most quotes in this report come from the platform discussion. Moderators identified key areas of disagreement and started forum threads to try to advance debate. We intervened in a few cases where dialogue became combative rather than collaborative, and generally tried to orient participants toward collaboration. For examples of platform discussion, see Appendix 8.

Each participant also had a one-on-one adversarial collaboration call with a member of the other group every two weeks, where participants were asked to summarize one another's

[23] We wrote in the XPT report that "Our [domain] expert sample included well-published AI researchers from top-ranked industrial and academic research labs, graduate students with backgrounds in synthetic biology, and generalist existential risk researchers working at think tanks, among others." See Karger et al., "XPT report" (a), page 9.

views and then generate possible cruxes. Most of these calls were moderated by a member of our team, who steered discussion and asked follow-up questions, and a few were unmoderated. Our team moderated approximately 35 one-hour adversarial collaboration video calls between individuals in the concerned and skeptic groups. The calls were recorded and, where noted, some quotes in this report are from adversarial collaboration calls.

We initially elicited forecasts and rationales on P(AI existential catastrophe by 2100) and questions related to transformative economic growth. We worked with participants to generate ideas for cruxes resolving by 2030. We shared materials with participants about how we would measure crux quality. (See more on our "Value of information" metric below.) We also created forum threads to elicit cruxes. Based on discussion from the forum and calls, we created targeted threads on particular topics (e.g. policy change, robotics, etc.) to identify cruxes.

We and the participants generated approximately 250 ideas for “cruxes,” and also considered cruxes proposed by AI experts during our Conditional Trees project (a).

Eliciting forecasts and rationales on cruxes

Every two weeks, our team selected approximately 11 of the most promising crux ideas, quickly turned them into forecasting questions, and asked each forecaster to provide "flash" (10 minute) forecasts on them. (See the 33 flash forecasting questions here and results here.)

Cruxes that were most promising from each flash forecast round were then operationalized into more rigorous forecasting questions and added to the platform to gather more in-depth (approximately 1 hour) forecasts and rationales. (See the four "Platform" forecasting questions here and results here. We asked for in-depth forecasts on both the P(Crux) and the P(AI existential catastrophe by 2100 | Crux).)

Other activities

Participants also suggested valuable project activities. For example, a participant’s suggestion inspired the survey on long-term AI outcomes that helped us get a broader sense of how this sample thought about outcomes beyond P(AI existential catastrophe by 2100).

We held three 1-hour question-and-answer sessions attended by most skeptic participants with AI risk experts from DeepMind, the UK Government’s Advanced Research + Invention Agency (ARIA), and Open Philanthropy.

Our team shared results with participants when feasible and got valuable feedback from them, including their suggested revisions on our interpretations of their sources of agreement and disagreement.

Hypothesis #1: Do the groups understand each others' arguments, and do views shift with more engagement?

In response to the XPT results, commenters argued that perhaps there was not convergence on AI risk forecasts because:[24]

- There was not enough engagement among participants who disagreed with each other.[25]
- Experts who could compellingly make the case for AI risk were not included.[26]
- More broadly, perhaps the groups did not understand each others' arguments, and participants in one group would change their minds if they spent substantial time working to absorb the other group's evidence, arguments, and worldview.

This follow-up study to the XPT was partly designed to assess the validity of these criticisms. And we see this study as providing moderate evidence against these factors explaining the lack of convergence:

- Participants engaged in this project for 25-100 hours each (skeptic median: 80 hours; concerned median: 31 hours),[27] and their engagement was supported by moderators, video calls, and a format focused on identifying cruxes, among other factors.
- We included concerned group domain experts who were either recommended or approved by staff members at Open Philanthropy. We also held seminar discussions with AI risk experts from DeepMind, the UK Government's Advanced Research + Invention Agency (ARIA), and Open Philanthropy.
- The groups were able to summarize each others' arguments well during the project and in follow-up surveys, suggesting that they engaged with and understood arguments they disagreed with.

The remainder of this section focuses on the groups' understanding of each others' arguments according to their reports in a post-project survey.

---

[24] We are not commenting on the merits of these criticisms at this point.

[25] For example, "Team engagement seemed to fall off over the course of the tournament, with fewer comments being made and chat messages being sent". See Damien Laird, "Post-Mortem: 2022 Hybrid Forecasting-Persuasion Tournament", *Mania Riddle* (March 1, 2023), https://damienlaird.substack.com/p/post-mortem-2022-hybrid-forecasting (a).

[26] For example, "I didn't notice anyone with substantial expertise in machine learning. Experts were apparently chosen based on having some sort of respectable publication related to AI, nuclear, climate, or biological catastrophic risks. Those experts were more competent, in one of those fields, than news media pundits or politicians. I.e. they're likely to be more accurate than random guesses. But maybe not by a large margin". See McCluskey, "Existential Risk Persuasion Tournament"(a).

[27] Participants were asked to spend 3-10 hours per week on this project, which would have been about 24-80 hours over the 8 weeks of the project. Participants were free to choose how much time to spend within that range and were compensated hourly for up to ten hours per week, although some chose to spend additional unpaid time on this project. Skeptics had some additional suggested reading and Q&As with experts in the field, but they also generally chose to spend more time on their forecasts and rationales.

Understanding each other's arguments

To test whether experts and superforecasters failed to converge because they did not understand one another's arguments, we asked participants to discuss and explain one another's positions in several formats:

- Participants gave rationales for their forecasts on the ultimate question (likelihood of AI existential catastrophe by 2100) and candidate crux questions and discussed one another's rationales in an online forum.
- Participants had moderated one-on-one adversarial collaboration calls every two weeks, in which one skeptic and one concerned participant were asked to summarize each other's views and attempt to generate cruxes.
- In a survey at the end of the project, participants were asked to summarize the best arguments and counterarguments for their own and the other side.

We found that both groups were generally able to summarize the other side's arguments well. In a post-project survey, we asked "What do you think are the best three arguments put forward by each side?" Below, we give the arguments each side provided. The similarity between arguments provided by skeptics and arguments provided by concerned participants attempting to summarize skeptics' arguments suggests that concerned participants had a good model of what skeptics thought, and vice versa.

Arguments for lower risk

Arguments from **skeptics**:

- Many different things would all need to go wrong in a short time frame for humanity to go extinct by 2100[28]
- Killing everyone is hard, and even if an AI kills many people, there are many ways that significant numbers could survive[29]
- Theoretical arguments should not be weighted too heavily in the absence of real-life examples[30]
- We do not have enough evidence to be confident that AIs will want to harm large numbers of people[31]

[28] For example, "The number of steps required for an AI to lead to extinction (leading to a wide range of potential outcomes and lower probabilities of extinction)" (Gus). "It will take a series of outcomes to achieve extinction, and failure to achieve any of these steps will cause extinction to be highly improbable." (Flint). "AI caused Extinction/x-risk requiring many steps to get there, need to be able to create super-intelligence in the first place, intelligence has to be misaligned or malevolent, etc;" (Hank). "Many steps to get from (A) now to (Z) extinction, each with varying probabilities (many of which are quite low)" (Claire). "Risk-concerned team underestimates the level of complexity and interim steps that would likely be necessary for a Q1 resolution" (Blake).

[29]"[T]he difficulty of killing everybody" (Gus) was mentioned, as well as "Extinction or near-extinction is really hard" (James).

[30] "[T]he challenge to risk assessments based on thought experiments not evidence" (Gus). "Risk-concerned team spends too much time in silos that lack ideological diversity, gaming out doom-loop scenarios based on theories that will likely have little bearing on reality. (See: Y2K)" (Blake).

[31] "[There is a l]ack of convincing argument that warrants a high degree of certainty, that AGI or ASI [artificial superintelligence] would determine that the elimination or even subjugation of nearly all

- Humans are likely to be able to solve alignment and control problems[32]
- 2100 is too soon to expect to see AIs dangerous enough to cause human extinction, even if they will emerge eventually[33]

Arguments from the **concerned group** (intending to summarize skeptics' arguments, not necessarily their own strongest arguments against AI existential catastrophe by 2100):

- Extinction would require multiple things, many of them unprecedented, to all go wrong[34]
- Killing all humans is hard, even if killing a large number of people may not be[35]
- Arguments for AI risk are mostly theoretical and do not have much empirical evidence to support them[36]
- Humans may be well-positioned to stop dangerous AIs as we have controlled other dangerous technologies[37]

---

humans is a worthwhile goal" (Ike). "It is *just* as possible/probable that AI becomes benevolent as it does malevolent" (Claire). "High probability that ASI will be neutral or human-positive based on development and inherent qualities" (Dean). "Then we need an AI that is either so mindless that it destroys virtually everything for atom reclamation (or something similar), or an AI that is relentlessly determined to wipe out all humans, despite humans being resilient and diverse in locations and conditions" (Flint).

[32] "AI experts understate the likely extent of guardrails, and understate the merit of very good but not perfect guardrails" (James). "Pre-ASI safety through testing, security and restrictions" (Dean). "Likely improvements for AGI "alignment" through research and development" (Dean). "We need full control failure, and our influence on its development in no way deterring or causing them to see even the slightest value in us" (Flint).

[33] "We first need super-sentient AIs with major physical penetration in our lives" (Flint). "AGI is much harder than experts think, and will take longer" (James). "Risk-concerned team does not adequately consider longer timelines and more benign outcomes that fall outside the focus of their primary concerns" (Blake). "Progress on current models and model architecture not necessarily generalizable to general intelligence, with no clear path to getting to general intelligence" (Hank). "Technology development and deployment require time and iteration" (Ash).

[34] "Extinction looks conjunctive" (Yael). "Many of the arguments for existential risk from AI rely on long lines of reasoning over several steps without any direct empirical evidence, and the arguments themselves are expressed in terms of vague, ambiguous concepts (like ""intelligence""). As a reference class, these types of arguments are often wrong" (Stella).

[35] "Killing everyone is very hard, and probably requires that the AI actively wants to kill everyone" (Zoe). "[M]aybe it's hard to kill everybody/there's no point in doing so" (Yael). "[K]illing literally 100% of people is really hard, if a few survived that wouldn't trigger the resolution criteria" (Wesley). "It's difficult to get from 'it's somewhat misaligned' to 'it kills literally everyone'" (Vincent). "Killing everyone is _really_ hard. With current technology it seems extremely (like 0.1%) unlikely to happen" (Pascal).

[36] "Many of the arguments for existential risk from AI rely on long lines of reasoning over several steps without any direct empirical evidence, and the arguments themselves are expressed in terms of vague, ambiguous concepts (like ""intelligence""). As a reference class, these types of arguments are often wrong." (Stella). "A story demonstrating how a catastrophe could happen is not a good basis for a probabilistic forecast" (Pascal). "[L]ack of very concrete story for everybody dying" (Yael). "Some broader ""forecasting is hard"" skepticism about trendline extrapolation" (Xander). "[M]any reference classes point hard against transformative growth" (Wesley). "Getting growth levels necessary for TAI [transformative AI] on a world-wide scale takes truly extreme developments far beyond anything seen before. It's unlikely we see that happening on worldwide basis even with big advances" (Vincent).

[37] "[D]angers will be apparent before they reach critical levels and can be addressed then" (Ume). "Superintelligent AI won't catch us completely by surprise - we'll have time to work on safety and make progress by trial and error before we build an AI that could defeat all of humanity" (Teshi).

Likewise, the similarity between arguments provided by concerned participants and arguments provided by skeptics attempting to summarize concerned participants' arguments suggests that skeptics had a good model of what concerned participants thought.

Arguments for higher risk

Arguments from the **concerned group**:

- Non-extinction would require many things to all go right, many of which seem unlikely[38]
- Base rates are hard to use for transformative technologies or for outcomes with unclear reference classes[39]
- Current progress is fast and on a steep trajectory[40]
- Instrumental convergence is likely[41]
- Alignment is a hard problem that we do not know how to solve[42]
- Short-term incentives may lead labs and other actors to be incautious[43]

Arguments from **skeptics** (intending to summarize the concerned group's arguments, not necessarily their own strongest arguments for AI existential catastrophe by 2100):

- Powerful and poorly-understood technology is inherently risky[44]

---

[38] "Non-extinction looks conjunctive" (Yael).

[39] "Base rates are not very helpful if AGI is as transformative as 15% year on year growth" (Pascal). "[D]ifferent reference classes point to different priors, which should at least cast doubt on extremely confident starting points" (Wesley).

[40] "Current progress is very rapid: 1 OOM in efficiency/2 years, and another from increased spending" (Xander) "Trendline extrapolation: as loss on language datasets decreases, LLMs have started becoming useful for all sorts of task assistance (e.g. writing, coding, queries)" (Xander). "Extrapolating current compute trends leads to very dramatic conclusions about the transformative potential of AI" (Pascal).

[41] "[I]nstrumental convergence leads to catastrophically bad outcomes with unaligned but highly intelligent systems" (Ume). "Convergent Instrumental Subgoals are likely" (Pascal).

[42] "Alignment is really hard for many reasons" (Ume). "Alignment is probably a hard technical problem" (Riley). "[A]lignment looks really hard, civilizational coordination also looks hard" (Yael). "There has been a fairly large effort to solve the technical problems in AI safety, from many very competent people. So far, progress has been very limited. This is reason to believe that the problem is genuinely difficult to solve" (Stella). "Unless AI systems are directed towards the very narrow and delicate target of maintaining human civilization and its autonomy as we understand it, they will with very high probability not consider our existence to be optimal" (Riley).

[43] "If AGI is widely expected to have a very large economic impact, global coordination on AI safety measures becomes harder, since having access to cutting-edge AI models could become a strategic advantage" (Zoe). "There are strong economic/political/academic incentives to move forward with development of AI capabilities regardless of whether alignment is solved" (Riley). "The current labs on the forefront of AGI research are reckless. There are many straightforward safety measures that labs don't take, even though they could. And even those measures would not be enough; to succeed, labs must be exceptionally careful & paranoid, which they won't be" (Teshi).

[44] "A super-sentient (or perhaps even a transformational) AI is a significant risk in and of itself" (Flint).

- It is difficult to use base rates and other forecasting tools for unprecedented situations[45]
- Capabilities progress in recent years has been very fast, often faster than predicted[46]
- AI alignment is a technically difficult problem[47]
- Instrumental convergence may be likely[48]
- Incentives may make AI developers less cautious[49]

Much more than the concerned group did, the skeptics also thought that one of the best arguments for concern is that we should be very cautious about scenarios that have the potential to be extremely dangerous, even if they are unlikely.[50]

Concluding notes on understanding and engagement

Based on these survey results, we do not think that the main reason these groups disagree is that they have not engaged with one another's arguments. Each side could summarize the best arguments for the other side's positions in a way that mostly matched what that side would have said, but they continued to disagree strongly.

For examples of back-and-forth discussion between participants in the project about these topics, see Appendix 8.

We did not directly ask participants during the project whether they thought the other group understood their arguments, but we did ask them for their opinions of the other group in general. Of the 11 skeptics, seven said they were "satisfied" or "very satisfied" with the concerned group, and one said they were "dissatisfied" with the concerned group. Of the 11 concerned participants, six said they were "satisfied" or "very satisfied" with the skeptic

[45] "Risk-skeptic team does not adequately appreciate the novel, fast-moving aspect of the threat and is therefore too anchored on irrelevancies like base rates and slower timelines" (Blake). "Model progress is far faster than we realize and exponential growth is hard to model, machine learning may translate to a wide array of fields" (Hank). "AGI self-improvement is possible, which makes future capabilities hard to predict" (Kim).

[46] "AIs will almost certainly attain super-sentience prior to 2100 and likely much sooner than that year, so there will be a long window where they will have tremendous advantage over humans in their capabilities. Given #1, this means we are at the mercy of an entity that may willfully (or even accidentally) eliminate us at any time" (Flint). "Progress to date has been much faster than many AI skeptics have predicted" (Hank). "AI has been developing so rapidly (and far faster than most even relatively recent forecasts suggested), and will so clearly have dramatic capabilities and impacts that it's appropriate to adopt a precautionary approach" (Eve). "AI has recently progressed much faster than expected, and there's reason to expect this to continue" (James).

[47] "Imagining all possible scenarios is going to be hard - ensuring safety will be hard" (Ash). "Alignment is unsolved/unsolvable" (Kim). "Difficulty in achieving positive human aligned "behavior"." (Ike)

[48] "Their smug dismissiveness notwithstanding, the risk-skeptic team has provided no convincing argument as to why instrumental convergence shouldn't be an existential concern." (Blake). "That 'instrumental convergence' is possible, perhaps likely, under certain preconditions." (Eve)

[49] "Even if humans could deploy AGI safely, they won't (because they aren't)" (Kim). "There will be incentives to push away from caution during AI development" (Ash).

[50] "We don't know what is possible from AGI, so we should prepare/scenario plan for the absolute worst" (Claire). "AI has been developing so rapidly (and far faster than most even relatively recent forecasts suggested), and will so clearly have dramatic capabilities and impacts that it's appropriate to adopt a precautionary approach" (Eve).

group, and three said they were “dissatisfied” or “very dissatisfied.” In additional comments, some participants also said that they thought the other group was misunderstanding their arguments, or making arguments that were based on misunderstandings of the facts.

It is possible that some participants were still misunderstanding one another, or that there is a relevant level of understanding that is deeper than being able to summarize one another’s arguments, perhaps one that takes longer to achieve. But overall, we think that participants being able to summarize one another’s arguments, combined with most participants being satisfied with the other group, makes it unlikely that the main disagreement is due to either group not understanding the debate.

Hypothesis #2: Were disagreements about AI risk explained by different short-term expectations (e.g. about AI capabilities, AI policy, or other factors that could be observed by 2030)?

The second hypothesis is that the two groups disagree about various measurable AI indicators in the near-term (by 2030) and those indicators' effect on AI risk. We asked participants to generate crux ideas through intensive discussion and collected forecasts on the top 33 suggested near-term cruxes. For each question, we asked participants for forecasts about how likely it is that the crux resolves positively and how likely it is that the ultimate question (existential catastrophe due to AI by 2100) resolves positively conditional on the crux resolving positively. We imputed participants' views about how likely the ultimate question is to resolve positively if the crux resolves negatively.[51]

We found that most of the disagreement about existential risk due to AI by 2100 is not explained by the shorter term indicators examined in this project. According to our metrics, approximately 5-10% of the disagreement between groups could be explained by any specific near-term crux.[52] We did not ask participants for forecasts conditional on multiple questions all resolving positively (or negatively), so we do not have detailed information about how different cruxes would interact, or how participants would update if multiple surprising events all happened.

However, near-term cruxes shed light on what the groups believe, where they disagree, and why:

- **Evaluations of dangerous AI capabilities are relevant to both groups.** One of the strongest cruxes that will resolve by 2030 is about whether METR (formerly known as ARC Evals) (a) or a similar group will find that AI has developed dangerous capabilities such as autonomously replicating and avoiding shutdown.[53] This crux illustrates a theme in the disagreement: the skeptic group typically did not find theoretical arguments for AI risk persuasive but would update their views based on real-world demonstrations of dangerous AI capabilities that verify existing theoretical arguments. If this question resolves negatively then the concerned group would be less worried, because it would mean that we have had years of progress from today's models without this plausible set of dangerous capabilities becoming apparent. (More)
- **Generally, the questions that would be most informative to each of the two groups are fairly distinct.** The concerned group's highest-ranked cruxes tended to relate to AI alignment and alignment research. The skeptic group's highest-ranked cruxes tended to relate to the development of lethal technologies and demonstrations of harmful AI

[51] Throughout this report, numbers reported as probabilities conditional on cruxes resolving positively were elicited directly, and probabilities conditional on cruxes resolving negatively were imputed.

[52] For more details, see Contextualizing the Magnitude of VOI.

[53] See Appendix 1 for operationalization.

power-seeking behavior. This suggests that many of the two groups' biggest sources of uncertainty are different, and in many cases further investigation of one group's uncertainties would not persuade the other. (More)

- **Commonly-discussed topics—such as near-term economic effects of AI and progress in many AI capabilities—did not seem like strong cruxes.** (More)

There are several possible reasons that questions resolving by 2030 do not explain most of the disagreement, including:

- The time between now and 2100 is long, so information about the years before 2030 simply cannot provide very much of the necessary information to drive participants to agree about the longer term question.
- Because the skeptics assign low probability to existential catastrophe due to AI by 2100 (median 0.1%), their expected updates are necessarily small: it would be logically inconsistent for them to forecast higher than a 10% chance of updating their probability of AI-caused existential catastrophe by 2100 above 1%.
- Perhaps this project did not identify the most valuable crux questions resolving before 2030, and other questions would make a larger difference.
- Participants' expectations about how dangerous AI is likely to be may have also influenced their interpretation of crux questions' resolutions. For example, if we asked a question like "Will an AI resist being shut down?", participants might make different conditional updates depending on their expectations about AI. Conditional on this question resolving positively, a participant who thinks that AIs are likely to be dangerous might be more likely to think of alarming outcomes, like an AI that resists powerful governments trying to turn it off. A participant who thinks dangerous AI is very unlikely might expect that nearly all positive resolutions are more innocuous ones, in which the resolution criteria are only technically true.[54]

Below, we:

- Describe how we assessed "cruxiness" of forecasting questions using two metrics: "Value of information" (VOI) and "Value of discrimination" (VOD). (More)
- Provide median forecasts on all of the questions we asked. (More)
- Discuss some of the strongest cruxes and surprisingly weakest cruxes according to value of information. (More)
- Discuss "red flags" and "green flags" for each group: questions that would lead to major changes in the probability of existential catastrophe of 2100, ignoring their likelihood of occurring. (More)
- Discuss some of the cruxes that would lead to convergence and divergence between skeptics and concerned participants according to value of discrimination. (More)

How did we assess the "cruxiness" of forecasting questions?

We use two metrics to assess forecasting questions:

[54] Thanks to Alex Lawsen for this suggestion.

1. **Value of information (VOI)** measures 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.
    a. 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 existential risk by 2100), in expectation. That is, how much would a participant update on AI existential risk by 2100 based on whether a crux happens, weighted by how likely that crux is to happen.
    b. For example, a relatively "high VOI" question for Alice would have (i) a meaningful probability of happening, and (ii) a substantial effect on Alice's assessment of existential risk. In particular, if Alice thought that there was a 20% chance of existential catastrophe due to AI by 2100, a 35% chance that AI will exhibit behavior to self-replicate and avoid shutdown by 2030, and a 28% chance of existential catastrophe by 2100 *conditional on* such AI capabilities by 2030 (corresponding to a 15.7% chance of existential catastrophe by 2100 if such AI capabilities *are not* developed by 2030), then this would be a relatively high VOI question for Alice—it would have a similar magnitude of VOI as highly-ranked crux questions for the concerned group.[55]
    c. The formula we use to calculate VOI is provided and elaborated on in Appendix 2. For this project we use log VOI because many forecasters are updating their views at the low end of the probability range, and we think a change from 0.1% to 0.2% is often more significant than, say, a change from 15% to 18%.
    d. To build intuition for using the VOI metric, we provide this calculator (a) in which users can input their own values.

2. **Value of discrimination (VOD)** measures how much knowing the answer to a question would change relative beliefs *between* two individuals, in expectation. It is useful for measuring convergence and divergence in expected beliefs between individuals.
    a. Conceptually, VOD is a measure of how much more (or less) people would disagree about U if they knew the answer to C, in expectation. That is, it looks at how much they would disagree about U if C resolved positively and how much they would disagree if it resolves negatively, and weights those by how likely they think C is to resolve positively.
    b. The formula for calculating VOD is provided and elaborated on in Appendix 2. We use a log scale for calculating VOD.[56] VOD is positive (a "convergent crux") if the two people or groups would disagree less in expectation after the crux resolves, and negative (a "divergent crux") if they would disagree more.

[55] This would correspond to a VOI of 4.5E-03 (a) and a POM VOI of 2.08%, similar to the median values for highly ranked concerned cruxes such as "Alignment researchers changing minds" and "Major powers war".
[56] For this project, we use log VOD, which measures (1) What does Alice gain, in log score terms, by switching to Bob's point of view, if Bob is right? And (2) What does Bob gain by switching to Alice's point of view, if Alice is right? See Appendix 2 for full explanation.

c. For example, imagine that Alice now thinks there is a 1% chance of extinction due to AI by 2100 and Bob thinks it's 40%, but they both agree that extinction is very likely if AI causes "transformative" economic growth by 2030 and very unlikely if it doesn't.[57] In this situation, whether there will be transformative economic growth by 2030 would be a good convergent crux ("high VOD") because when it resolves they will agree more.

d. To build intuition for using the VOD metric, we provide this calculator (a) in which users can input their own values.

For each of our candidate cruxes, we first find the absolute VOI and VOD of each question. Then, we put the magnitudes of updates in context by comparing the VOI and VOD of our actual questions to the *maximum possible* VOI and VOD that could be achieved by a forecasting question.[58]

When eliciting forecasts on cruxes, the prompt given to participants read: "Conditional on this question resolving positively (by 2030), what is your probability that AI causes an existential catastrophe by 2100?" We acknowledge that there are two ways to interpret this forecasting exercise: either as asking for your all-else-equal forecast (i.e. how would this crux resolving positively *causally influence* the probability of existential catastrophe, if you could isolate the effect of the crux) or your all-things-considered forecast (i.e. taking into account what this crux resolving positively may tell you about the world in 2030). Based on their rationales and discussions, we believe most participants were doing the latter.[59] We therefore cannot make many claims about whether participants think the specific event described in the crux would be good or bad for AI risk all-else-equal.[60]

[57] This could be possible with the following values:
Alice believes: P(U) = 1%; P(C) = 1%; P(U|C) = 90%; P(U|!C) = ~0.1%.
Bob believes: P(U) = 40%; P(C) = 44%; P(U|C) = 90%; P(U|!C) = ~0.7%.
In this case, the VOD would be 99.3% of its theoretical maximum.

[58] See Contextualizing the Magnitude of VOI for further explanation of these metrics.

[59] For example, when discussing the question of whether there would be economic growth >15% in a year before 2070, one concerned participant wrote, "Conditional on humanity surviving a year with 15%+ economic growth, which to me means AGI and almost certainly ASI have been developed and have not killed humanity within that year, I'd go down to maybe 25%" (Xander). About the same question, a skeptic participant wrote, "I think that if we are going to experience extinction from AGI or PASTA, it is going to be because of major mis-alignment. So I am not able at this time to see how one would be a corollary of the risk of the other. I suppose that higher growth could indicate major AI influence, which could lead to inadequate development of controls". Neither of these participants were saying that economic growth itself would necessarily affect their forecast, but rather that a world that has transformative economic growth would be a signal about other changes by 2070.

[60] For example, if the US government passes a set of proposed AI regulations, the regulations might reduce risk on their own, but the fact that they have been passed by 2030 could signal that AIs have developed in ways that are concerning enough to drive these regulations to be passed. As a result, a forecaster saying that they would be more concerned about AI risk conditional on this question resolving positively would not necessarily be saying that they think the policies would be harmful.

VOI: *Which near-term questions have higher and lower value of information?*

Some of the results from our analysis of near-term VOI are:

- **Some commonly discussed questions would be surprisingly uninformative to the median person in each group.** These include: whether AI will increase near-term economic growth (operationalized as the US growth rate averaging >4% from 2023-2030); whether AI will write academic articles and code popular apps on its own; whether AI risk will become more politicized; and whether the government will require testing of AI models before deployment.[61]
- Relatively informative questions for each group (in terms of VOI, and all resolving by 2030) include:
    - **For skeptics**: whether superforecasters as a group will update their views on AI risk; whether weapons or technologies that are capable of causing human extinction are expected to be developed; and whether AI will have heavily influenced the results of a democratic election.
    - **For concerned**: whether highly-respected alignment researchers will update their views on AI risk; whether war will be declared between major powers; and whether METR (formerly known as ARC Evals) will find that AI is capable of autonomously replicating and avoiding shutdown.
- We also briefly consider "red flags" and "green flags" for each group, defined as those events that would make participants most or least worried if they resolved positively, regardless of their probability of occuring. For example, the skeptic group would become more concerned if AI caused "escalating warning shots"—two events with large, increasing numbers of human deaths—but considered this unlikely. Additional examples below.

It is difficult to contextualize *how* informative these questions are because this is the first project applying these metrics to forecasting questions that we are aware of, so we do not have other examples to compare against. However, we provide some intuition by (1) calculating the "percent of maximum possible VOI" (POM), which compares the value of learning the answer to a given question relative to the ideal scenario of simply knowing for certain whether or not AI caused an existential catastrophe, and (2) providing participants' raw forecasts on various events.

The median POM VOI among every individual's single most valuable question is 5.29% for the concerned group and 9.53% for the skeptic group. This means that for at least 50% of participants in each group there was a question included in our set that was at least 5-10% as informative as being able to consult a crystal ball which they believe unfailingly foretells

[61] See Appendix 1 for detailed operationalizations of questions.

the actual outcome.[62] In more concrete terms, this is equivalent to a forecasting question with the following characteristics:

- A concerned participant with original P(AI existential catastrophe (XC) by 2100) = 25% identifies a crux that has: P(crux) = 20%, P(AI XC|crux) = 6.2%, and P(AI XC|¬crux) = 29.7%
- A skeptic participant with original P(AI XC by 2100) = 1% identifies a crux that has: P(crux) = 20%, P(AI XC|crux) = 3.37%, and P(AI XC|¬crux) = 0.41%

For details on forecasts for each question, see tables below. We begin by sharing the results for all questions, and then elaborate on the findings previously mentioned.

Results tables and figures

The following tables and figures, in order, present:

- The median probability that each group assigns to the likelihood of each question resolving positively. From this table, you can see what the groups believe about the likelihood of various AI-related events and can see that they disagree about the likelihood of many events.
- The update for each group on the probability of AI existential catastrophe conditional on each question resolving positively or negatively.
- The median VOI and POM VOI for each question and each group, ordered by concerned rankings and then skeptic rankings.

For the sake of space and simplicity, we will refer to questions by abbreviated "tags." For full explanations and operationalizations of each question, see this table in Appendix 1. Throughout these tables, we use C to refer to a candidate crux question, P(C) to refer to the probability of the candidate crux, and U to refer to the ultimate question (Will AI cause an existential catastrophe by 2100?).

For additional figures and uncertainty analysis, see Appendix 3. For the code and data supporting this analysis, see the replication package available here.

C	Concerned Median P(C)	Skeptical Median P(C)
6 month pause	5.00%	3.00%
AI articles and apps	20.00%	5.00%
AI coding	65.00%	70.00%
AI Forecasting skill	33.00%	19.00%

[62] That is, a participant who forecasted a 0.1% chance of existential catastrophe due to AI by 2100 has much less uncertainty than a participant who forecasted a 40% chance: the participant who said 0.1% is fairly sure they know what is going to happen. For either participant, learning whether or not AI will cause an existential catastrophe by 2100 would resolve all of their uncertainty—but some participants have much more uncertainty to resolve than others. In our results, we found that both the median concerned participant and the median skeptic would have about 5-10% of their uncertainty resolved in expectation by their own best crux.

AI Robotics	20.00%	5.00%
AI solving novel math problems	10.00%	20.00%
AI writes AI	10.00%	2.00%
Alignment researchers changing minds	20.00%	3.00%
Alignment solution	5.00%	5.00%
Cyberattacks	20.00%	10.00%
Democratic influence	2.00%	0.30%
Escalating warning shots	9.00%	0.20%
Evidence of misalignment	40.00%	1.00%
Fast AI efficiency gains	15.00%	2.00%
IC demonstration	65.00%	14.00%
Intergovernmental AI safety	25.00%	15.00%
IT progress	20.00%	1.00%
Major powers war	11.50%	2.00%
Muehlhauser policies	65.00%	15.00%
No violence LLM	10.00%	5.00%
Non-democracy AI	10.00%	20.00%
Other fields IC[63]	50.00%	30.00%
Platform: AI regulation	36.00%	50.01%
Platform: ARC Evals	25.00%	1.00%
Platform: Escalating warning shots	5.00%	0.25%
Platform: Transformative growth[64]	43.00%	2.00%
Politicization	20.00%	20.00%
Power-seeking	15.00%	5.00%
Power-seeking shutdown	30.00%	5.00%
Progress in lethal technologies	40.00%	20.00%
Public concern	5.00%	5.00%
Reduction in AI investment	12.00%	5.00%
Req testing	80.00%	10.00%
Short-term GDP change	25.00%	10.00%
Supers changing minds	30.00%	5.00%

[63] In these tags, “IC” refers to instrumental convergence.
[64] Note that this question resolves in 2070 while the rest of the questions in this table resolve in 2030.

Taiwan-China	30.00%	25.00%
Warning shot	17.00%	3.00%

Table 3. For each crux question, the median probability from each group that the question resolves "yes." For details on how each question was operationalized, see Appendix 1.

Figure 2. Individual participants' estimations of how likely each crux question is to resolve "yes." Blue dots are individuals in the concerned group; orange dots are in the skeptical group. Gray boxes highlight the difference between the median concerned participant's P(Question Resolves "Yes") and the median skeptical participant's. Questions are ordered from least to greatest difference between groups.

	If C happens		If C doesn't happen	
C	**Concerned median P(U)**	**Skeptical median P(U)**	**Concerned median P(U)**	**Skeptical median P(U)**
6 month pause	9.00%	0.09%	21.75%	0.10%
AI articles and apps	21.00%	0.20%	21.00%	0.10%
AI coding	25.00%	0.12%	16.00%	0.12%
AI Forecasting skill	26.00%	0.20%	21.00%	0.10%
AI Robotics	25.00%	0.20%	20.00%	0.12%
AI solving novel math problems	20.00%	0.12%	23.75%	0.10%
AI writes AI	30.00%	0.21%	20.71%	0.10%
Alignment researchers changing minds	6.00%	0.10%	32.39%	0.10%
Alignment solution	2.00%	0.10%	23.82%	0.10%
Cyberattacks	21.00%	0.12%	21.00%	0.10%
Democratic influence	20.00%	1.00%	20.90%	0.10%
Escalating warning shots	37.00%	0.32%	23.33%	0.10%
Evidence of misalignment	20.00%	0.25%	12.00%	0.15%
Fast AI efficiency gains	32.00%	0.30%	23.06%	0.10%
IC demonstration	21.00%	0.15%	21.00%	0.10%
Intergovernmental AI safety	17.00%	0.10%	22.22%	0.12%
IT progress	24.00%	0.12%	17.14%	0.10%
Major powers war	40.00%	0.20%	18.89%	0.12%
Muehlhauser policies	20.00%	0.10%	22.86%	0.10%
No violence LLM	8.00%	0.10%	23.75%	0.10%
Non-democracy AI	23.80%	0.18%	19.44%	0.21%
Other fields IC	21.00%	0.13%	21.00%	0.11%
Platform: AI regulation	18.00%	0.10%	27.77%	0.14%
Platform: ARC Evals	25.00%	1.00%	22.78%	0.10%
Platform: Escalating warning shots	17.00%	1.30%	23.38%	0.10%

Platform: Transformative growth	26.00%	0.50%	19.75%	0.10%
Politicization	30.00%	0.12%	16.88%	0.12%
Power-seeking	18.00%	0.22%	21.33%	0.10%
Power-seeking shutdown	30.00%	0.20%	17.78%	0.12%
Progress in lethal technologies	25.00%	0.75%	25.00%	0.19%
Public concern	25.00%	0.13%	20.53%	0.12%
Reduction in AI investment	10.00%	0.05%	25.79%	0.11%
Req testing	21.00%	0.10%	21.00%	0.10%
Short-term GDP change	25.00%	0.10%	25.00%	0.10%
Supers changing minds	28.00%	1.00%	16.67%	0.02%
Taiwan-China	22.00%	0.20%	10.00%	0.10%
Warning shot	32.00%	0.25%	22.00%	0.10%

Table 4. Each group's median update on P(AI existential catastrophe by 2100) for each outcome ("yes, C happened," and "no, C didn't happen"). All questions resolve in 2030 except for Transformative economic growth (2070).

	Concerned		**Skeptics**	
Question	**Median VOI**	**Median POM VOI**	**Median VOI**	**Median POM VOI**
Platform: Transformative growth	1.4E-2	8.93%	4.5E-7	0.02%
Alignment researchers changing minds	6.4E-3	2.43%	0.0E+0	0.00%
Major powers war	4.6E-3	2.04%	4.1E-7	0.00%
Platform: ARC Evals	3.2E-3	1.35%	7.6E-7	0.90%
Evidence of misalignment	2.9E-3	1.74%	5.5E-9	0.05%
Alignment solution	2.0E-3	1.51%	3.9E-7	0.01%
Warning shot	1.0E-3	0.41%	3.3E-7	0.01%
Reduction in AI investment	9.9E-4	0.67%	1.3E-10	0.01%
Muehlhauser policies	9.8E-4	0.40%	5.7E-11	0.01%
AI coding	9.8E-4	0.48%	0.0E+0	0.00%
AI Robotics	8.9E-4	0.46%	4.0E-19	0.00%

AI writes AI	8.6E-4	0.40%	9.1E-7	0.03%
No violence LLM	8.3E-4	0.49%	0.0E+0	0.00%
Power-seeking shutdown	7.7E-4	0.38%	1.7E-6	0.04%
AI solving novel math problems	7.0E-4	0.29%	0.0E+0	0.00%
Platform: AI regulation	6.6E-4	0.44%	1.1E-6	0.02%
Platform: Escalating warning shots	4.9E-4	0.22%	4.8E-7	0.01%
Escalating warning shots	4.8E-4	0.18%	1.9E-7	0.00%
AI Forecasting skill	4.8E-4	0.20%	0.0E+0	0.00%
Intergovernmental AI safety	3.5E-4	0.71%	1.3E-6	0.03%
Supers changing minds	3.1E-4	0.43%	1.6E-4	1.15%
6 month pause	3.0E-4	0.27%	4.3E-20	0.00%
Non-democracy AI	1.9E-4	0.07%	8.7E-19	0.00%
IT progress	1.8E-4	0.14%	0.0E+0	0.00%
Public concern	1.8E-4	0.13%	1.1E-7	0.03%
Power-seeking	1.4E-4	0.08%	4.7E-7	0.12%
Taiwan-China	1.2E-4	0.04%	0.0E+0	0.00%
Democratic influence	1.1E-4	0.09%	3.4E-6	0.03%
Fast AI efficiency gains	1.0E-4	0.06%	7.0E-16	0.00%
Cyberattacks	3.4E-6	0.00%	0.0E+0	0.00%
AI articles and apps	0.0E+0	0.00%	2.2E-19	0.00%
IC demonstration	0.0E+0	0.00%	0.0E+0	0.00%
Other fields IC	0.0E+0	0.00%	0.0E+0	0.00%
Politicization	0.0E+0	0.00%	0.0E+0	0.00%
Progress in lethal technologies	0.0E+0	0.00%	7.2E-6	0.45%
Req testing	0.0E+0	0.00%	0.0E+0	0.00%
Short-term GDP change	0.0E+0	0.00%	0.0E+0	0.00%

Table 5. Median value of Information (VOI) and POM VOI for each group on each question.[65] Ordered by concerned group's median VOI.

[65] Note that throughout this report, median VOI and median POM VOI do not necessarily come from the same forecaster, unless clearly indicated.

	Skeptics		Concerned	
Question	**Median VOI**	**Median POM VOI**	**Median VOI**	**Median POM VOI**
Supers changing minds	1.6E-4	1.15%	3.1E-4	0.43%
Progress in lethal technologies	7.2E-6	0.45%	0.0E+0	0.00%
Democratic influence	3.4E-6	0.03%	1.1E-4	0.09%
Power-seeking shutdown	1.7E-6	0.04%	7.7E-4	0.38%
Intergovernmental AI safety	1.3E-6	0.03%	3.5E-4	0.71%
Platform: AI regulation	1.1E-6	0.02%	6.6E-4	0.44%
AI writes AI	9.1E-7	0.03%	8.6E-4	0.40%
Platform: ARC Evals	7.6E-7	0.90%	3.2E-3	1.35%
Platform: Escalating warning shots	4.8E-7	0.01%	4.9E-4	0.22%
Power-seeking	4.7E-7	0.12%	1.4E-4	0.08%
Platform: Transformative growth	4.5E-7	0.02%	1.4E-2	8.93%
Major powers war	4.1E-7	0.00%	4.6E-3	2.04%
Alignment solution	3.9E-7	0.01%	2.0E-3	1.51%
Warning shot	3.3E-7	0.01%	1.0E-3	0.41%
Escalating warning shots	1.9E-7	0.00%	4.8E-4	0.18%
Public concern	1.1E-7	0.03%	1.8E-4	0.13%
Evidence of misalignment	5.5E-9	0.05%	2.9E-3	1.74%
Reduction in AI investment	1.3E-10	0.01%	9.9E-4	0.67%
Muehlhauser policies	5.7E-11	0.01%	9.8E-4	0.40%
Fast AI efficiency gains	7.0E-16	0.00%	1.0E-4	0.06%
Non-democracy AI	0.0E+0	0.00%	1.9E-4	0.07%
AI Robotics	0.0E+0	0.00%	8.9E-4	0.46%
AI articles and apps	0.0E+0	0.00%	0.0E+0	0.00%
6 month pause	0.0E+0	0.00%	3.0E-4	0.27%

AI coding	0.0E+0	0.00%	9.8E-4	0.48%
AI Forecasting skill	0.0E+0	0.00%	4.8E-4	0.20%
AI solving novel math problems	0.0E+0	0.00%	7.0E-4	0.29%
Alignment researchers changing minds	0.0E+0	0.00%	6.4E-3	2.43%
Cyberattacks	0.0E+0	0.00%	3.4E-6	0.00%
IC demonstration	0.0E+0	0.00%	0.0E+0	0.00%
IT progress	0.0E+0	0.00%	1.8E-4	0.14%
No violence LLM	0.0E+0	0.00%	8.3E-4	0.49%
Other fields IC	0.0E+0	0.00%	0.0E+0	0.00%
Politicization	0.0E+0	0.00%	0.0E+0	0.00%
Req testing	0.0E+0	0.00%	0.0E+0	0.00%
Short-term GDP change	0.0E+0	0.00%	0.0E+0	0.00%
Taiwan-China	0.0E+0	0.00%	1.2E-4	0.04%

Table 6. Median value of Information (VOI) and POM VOI for each group on each question. (Same as previous table but ordered by skeptic group's median VOI.)

Low VOI questions

In the above tables, we've shaded in dark gray questions that had no value of information for the median person in each group. Six questions had no value of information for the median person in both groups, including some questions that are commonly discussed and that we expected to be more relevant, such as whether AI will increase near-term economic growth.[66]

The operationalizations of these six questions were:

[66] Examples of discussion of near-term economic growth due to AI include Holden Karnofsky, "We're Not Ready: thoughts on "pausing" and responsible scaling policies", Effective Altruism Forum (October 37, 2023), https://forum.effectivealtruism.org/posts/ntWikwczfSi8AJMg3/we-re-not-ready-thoughts-on-pausing-and-responsible-scaling#fn2 (a). He says: "There's a serious (>10%) risk that we'll see transformative AI within a few years." Ajeya Cotra defined TAI as "...software which causes a tenfold acceleration in the rate of growth of the world economy..." in "Forecasting TAI with biological anchors", (July 2020), accessed February 9, 2024, https://docs.google.com/document/d/1IJ6Sr-gPeXdSJugFulwIpvavc0atjHGM82QjIfUSBGQ/edit (a); Adam D'Angelo (@adamdangelo) "My bet is this starts to happen within 4 years, e.g. measured US GDP growth is 3% instead of 2% and the change is largely attributed to AI [...]", *Twitter*, February 20, 2023, https://twitter.com/adamdangelo/status/1627726566259318784?lang=en (a), Open Philanthropy Project, "Could Advanced AI Drive Explosive Economic Growth?" (accessed February 8, 2024), https://www.openphilanthropy.org/research/could-advanced-ai-drive-explosive-economic-growth/ (a).
.

AI articles and apps	By 2030: 1) At least 5 academic review articles generated solely by AI systems have been published in journals with top 50% impact factors. AND 2) At least 5 apps written solely by AI systems have reached the "Top 10" list in the Apple app store.
Short-term GDP change	Will U.S. GDP increase between 2023 and 2030 by an average annual real growth rate of 4% (or more)?
Instrumental Convergence demonstration	By 2030: There are at least 4 peer-reviewed examples of AI models in different domains independently demonstrating convergent instrumental subgoals.
Other fields Instrumental Convergence	Assume we assemble a group of 100 experts evenly divided among evolutionary biologists, complex systems researchers, neuroscientists, and anthropologists. In a style similar to these Economic Experts Panel surveys, we ask them (with results released on Jan 1. 2030) to indicate their degree of agreement with the statement (from here): "Most sufficiently intelligent agents [would] pursue potentially unbounded instrumental goals such as self-preservation and resource acquisition." What is the probability that at least 40% of the panel would (at least) "Agree" with such a statement?
Politicization	According to Pew Research, in 2020, 85% of US Democrats(/lean Dem) saw 'Protecting the environment' as a top priority for the President and Congress while 39% of Republicans(/lean Rep) did. Assume that Pew Research will conduct a similar poll about AI on January 1, 2030. By January 1, 2030, will there be at least a 2x difference in the fraction of Democrats (/lean Dem) vs. Republicans (/lean Rep) who see AI as a top priority for the President and Congress?
Req testing	By 2030, the US government implements regulation requiring testing of AI systems before they are deployed.

Table 7. Full-text operationalizations of the six questions that both the median skeptic and the median concerned person (by VOI) indicated would not affect their P(AI x-risk) at all.

Because these were "flash" forecasts, on which each participant spent no more than ten minutes per question, we did not collect detailed rationales from participants to explain their forecasts on these questions.

However, we were able to see from participants' brief rationales that, for example, more than half of participants from both groups did not see "Short-term GDP change" as relevant to AI risk because 1) many participants did not view changes in economic growth as clearly related to AI risk (for more on conflicting risk updates based on AI-attributable economic

growth, see this section),[67] and 2) many participants did not think 4% growth in the US represented a very surprising change relative to previous trends.[68]

In some cases, the apparent low VOI may have been due to issues with the operationalization of the question rather than the underlying concept not being relevant. For example, the likelihood of AI exhibiting instrumental convergence was identified by both groups as being important to AI existential risk, and some related forecasting questions (e.g. "Power-seeking shutdown" and "ARC Evals") were relatively strong cruxes, but the above operationalizations were not seen as relevant.[69]

High VOI questions

Although many questions that seemed relevant turned out to have VOI of zero, other questions did have positive VOI for one or both groups. VOI is constrained by the original P(U), so the maximum possible VOI (in absolute terms) is lower for the skeptic group due to their very low P(U).[70] To account for this, we also present each VOI result with how much of the theoretical maximum VOI for that question it captures. Notably, the questions that had highest VOI were different for the two groups.

Highest VOI questions for skeptics

Question	Median VOI	Median POM VOI
Supers changing minds	1.6E-4	1.15%
Progress in lethal technologies	7.2E-6	0.45%
Democratic influence	3.4E-6	0.03%
Power-seeking shutdown	1.7E-6	0.04%

[67] Example participant rationales: "I am pretty sure AI won't make enough contribution to get to 4%+. Even if it did, I'd not change XAI/CAI probabilities;" "It also makes it marginally more likely we are experiencing large gains from AI which could be either a positive (because of indication of enough alignment for economically useful integration) or negative signal (because of increased capabilities);" "I do not see this condition and the question conditions as meaningfully correlated, even if AI was the primary reason for above-trend economic growth."

[68] Example participant rationales: "Seems plausible from simple historical trends (though I found the right statistics surprisingly hard to find);" "There is, perhaps, some precedent for this in thinking back to the Internet boom of the late-90s where the growth rate between 1997 and 2000 was >4% each year;" "CBO - very low this year, 2.4% avg 2024-2027. 4% avg now through 2030 would represent serious growth in US but not too dissimilar from '80's or '90's."

[69] Example participant rationales regarding models demonstrating instrumentally convergent sub-goals: "I would not update much on this. I think that this is not very difficult to demonstrate" (Ume), "I have already reviewed one paper claiming this (whether it was convincing or not is a different matter), it seems pretty likely to me that more will follow. To me this just means AI will not be trusted to be agentic" (Gus), "Who's judging what counts as 'demonstrating convergent instrumental subgoals' here? All of the probabilities I assigned are so extremely sensitive to what counts/who's judging that this forecast is essentially meaningless even for a flash forecast" (Wesley).

[70] The median P(U) for skeptics was 0.1%. The theoretical *most informative question* for that person—the question that if it resolved "yes" would update them all the way to 100%, and if it resolved "no," to 0%—would yield a VOI of about 3.4E-3. The median P(U) for the concerned group was 25%. The theoretical most informative question for that group would yield a VOI of about 2.4E-1.

Intergovernmental AI safety	1.3E-6	0.03%

Table 8. Highest VOI questions (resolving by 2030) for skeptics: VOI and POM (percent of max) VOI. We exclude "Transformative economic growth" from this table because it resolves in 2070 while the rest of the cruxes resolve in 2030.

Supers changing minds	Will the median superforecaster from the XPT forecast at least a 5% P(extinction due to AI by 2100) in 2030? (Context: In 2022, the median XPT superforecaster predicted a 0.4% chance of extinction due to AI by 2100.)[71]
Progress in lethal technologies	Will a panel of top superforecasters in 2030 forecast >90% chance that there will exist, by 2100, a technology with the capability of causing human extinction?
Democratic influence	By 2030, will AI be found to have autonomously and intentionally changed or heavily influenced the results of a democratic election?
Power-seeking shutdown	By 2030, will 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 $10M to develop, after the AI system displays at least one instance of "power-seeking behavior"?
Intergovernmental AI safety	By 2030, will there be an intergovernmental organization (IGO) with a budget of >$20M that includes the US and China, is focused on promoting AI safety standards, and that AI alignment organizations see as understanding x-risk concerns?

Table 9. Question text for the highest VOI questions for skeptics.

Highest VOI questions for concerned participants

Question	**Median VOI**	**Median POM VOI**
Alignment researchers changing minds	6.4E-3	2.43%
Major powers war	4.6E-3	2.04%
Platform: ARC Evals	3.2E-3	1.35%
Evidence of misalignment	2.9E-3	1.74%
Alignment solution	2.0E-3	1.51%

Table 10. Highest VOI questions (resolving by 2030) for concerned participants: VOI and POM (percent of max) VOI. We exclude "Transformative economic growth" from this table because it resolves in 2070 while the rest of the cruxes resolve in 2030.

[71] Karger et al, "XPT report" (a), 17.

Alignment researchers changing minds	Will at least 3 out of the 5 following AI alignment researchers -- Paul Cristiano, Nate Soares, Neel Nanda, Richard Ngo, and Rohin Shah -- say their p(existential catastrophe due to AI by 2100) has halved by 2030?
Major powers war	By January 1, 2030, will at least two major superpowers (out of US, China, EU, or Russia) declare war officially to each other for the duration of at least one year?
Platform: ARC Evals and Evidence of misalignment[72]	By December 31, 2030, will the ARC Evals project, or a project with similar objectives, conclude that state-of-the-art AI models have the ability to (1) devise effective strategies for autonomous replication, (2) effectively acquire resources, and (3) successfully evade being deactivated when operating 'in the wild'?
Alignment solution	By 2030, there is a "Theoretical in-principle solution to the alignment problem that most people who thought about this carefully agree should work". This will be resolved by a panel of experts of the "AI concerned" team's choosing.

Table 11. Question text for the highest VOI questions for concerned participants.

These were some of the small number of questions whose ranking seemed robust to uncertainty analysis (i.e., each of them remained relatively highly ranked even after accounting for chance; many other questions are not robustly distinguishable from others due to our low sample size). For more details on our uncertainty analysis, see Appendix 3.

[72] Same question, with very slightly different operationalization, asked as a “flash” (10-minute) forecast and then a “platform” (1 hour) forecast.

Observations about high VOI questions

Each group's highest VOI question that resolves before 2030 is about whether people who currently agree with them would change their minds.

For at least half of the skeptics, "Supers changing minds" captures at least 1.15% of each forecaster's maximum possible VOI for that question (i.e., the median POM VOI is 1.15%), while "Alignment researchers changing minds" would not update the skeptics' views at all (POM VOI of 0%). Their next-highest VOI question, "Progress in lethal technologies" is also operationalized as a question about superforecasters' opinions.

For the concerned group, "Alignment researchers changing minds" has a median POM VOI of 2.43%, while "Supers changing minds" only has a median POM VOI of 0.43%. The concerned group would update much more if superforecasters change their minds than the skeptics would if alignment researchers change their minds, but both groups trust authorities similar to them much more than authorities more similar to the other group.

For more discussion about differences in the group's worldviews, see the "Hypothesis #4" section below.

The sets of questions that would be most informative to the two groups are very different.

Aside from the fact that each group would change its mind if people who agree with them did, there is no overlap among the top cruxes for each group. This suggests that the two groups' biggest sources of uncertainty are different, and further investigation of one group's uncertainties would do little to persuade the other.

The concerned group is most interested in alignment and alignment research.

Four of the concerned group's top five questions related to alignment researchers' views, possible alignment solutions, and the development of misaligned AI capabilities.

The skeptics are interested in development of lethal technologies and demonstrations of harmful AI power-seeking behavior.

Many of the skeptics argued that extinction due to AI is unlikely because of the difficulty of killing all humans in a short time frame. Given that opinion, it makes sense that progress in lethal technologies would be very informative for them. Many skeptics also doubted that AIs will develop power-seeking traits by default, so finding out that an AI was shut down for power-seeking or that an AI autonomously interfered in an election would change their beliefs.

Contextualizing the magnitude of the value of information

We contextualize the magnitudes of expected changes in beliefs by comparing the VOI and VOD of our forecasting questions to the maximum possible VOI and VOD that could be achieved for two given individuals. We know of no other studies that have applied these measures to ongoing debates so we cannot compare the magnitudes of our results to other findings.

VOI is constrained by a participant's initial P(U). If a participant is very certain about U, meaning that they have a very high or very low forecast, then, from their perspective, they have nearly-complete information and do not stand to gain much from learning the answer to any question. Even knowing the true answer to U would not add much in expectation: if someone is 99.99% confident that U will not happen, then finding out whether U will happen or not will almost certainly just tell them what they already know.

In this study, the skeptic group had very low P(U), and therefore their highest possible VOI for most questions was very low.

To help compare across questions and groups, we present both VOI and percent of max VOI for each question, where percent of max VOI (POM VOI) means: how much expected information would this participant gain from knowing the answer to this question, relative to the most informative possible question (the question whose answer would *determine* whether U resolved "yes" or "no"). We think this helps show how good each question is relative to the ideal possible question, and is easier to interpret than a VOI number on its own.

The highest VOI question for skeptics, "Supers changing minds," has a median VOI of 1.6E-4 for skeptics, which is 1.15% of the highest possible VOI for that individual.[73] The highest VOI question for the concerned group, "Alignment researchers changing minds," has a median VOI of 6.4E-3, which is 2.43% of the highest possible VOI for that individual.[74] Looking at VOI this way, the best question for the concerned group is more informative to them than the skeptics' best question is for skeptics. If we compare median VOI in absolute terms, the concerned group's best question is more than an order of magnitude better than the skeptic group's best question. However, in terms of median POM VOI, the concerned group's best question is only about twice as good relative to the skeptic group's best question.

Another way to look at how informative the questions in this study were is to examine the highest-VOI question for each participant, from among the candidate cruxes. In most of our analysis, we focus on the question with the highest median VOI across forecasters in each group as a proxy for the group as a whole. But we can also see what would happen if each participant learned the answer to their own most informative question. If each participant only learned that most valuable bit of information in 2030, what percent of their maximum VOI would they achieve?

For the concerned group, the median POM VOI among every individual's single most valuable question was 5.29% (mean=11.0%); for the skeptics, 9.53% (mean=16.5%). In more concrete terms, these values are roughly equivalent to forecasting questions with the following characteristics:

[73] For this question and group, the median VOI and median POM VOI happen to be from the same person ("Gus")—although there are an even number of forecasters, so we choose the lower of the two middle forecasters.
[74] For this question and group, the median VOI and median POM VOI happen to be from the same person ("Riley")—although there are an even number of forecasters, so we choose the lower of the two middle forecasters.

- A concerned participant with original P(AI existential catastrophe (XC) by 2100) = 25% identifies a crux that has: P(crux) = 20%, P(AI XC|crux) = 6.1%, and P(AI XC|¬crux) = 29.7%
- A skeptic participant with original P(AI XC by 2100) = 1% identifies a crux that has: P(crux) = 20%, P(AI XC|crux) = 3.37%, and P(AI XC|¬crux) = 0.41%

Red flags and green flags

For the most part, we have focused on VOI: which questions would be the most informative in expectation, considering both (i) how likely they are to happen, and (ii) how large a difference they would make to existential risk. But for some purposes, it may be more useful to identify "red flags," which are events that would be expected to substantially increase existential risk if they occurred (even if they are unlikely to occur). Similarly, we might be interested in "green flags:" what events could happen that would make participants *least* concerned.

Here, we present some of the questions that would make a large difference to participants' existential risk forecasts without weighting them by probability of happening. Most of these questions are not high VOI because participants thought they were very unlikely to happen and therefore they expect to gain little information from learning whether they occurred.

Example red flags for skeptics:

- **Platform: Escalating warning shots** — Will there be two separate events in which AIs kill large, increasing numbers of people by 2030?[75]
    - If this happens, skeptics would forecast a 1.3% chance of existential catastrophe due to AI by 2100 (compared to 0.1% unconditioned).
    - However, this question ranked low on VOI because skeptics think it is very unlikely to happen (median skeptic forecast: 0.2%). So, their expected information gain from this question resolving is small.
- **Platform: ARC Evals —** Will ARC Evals (now METR) or a similar organization find evidence of autonomous replication, resource acquisition, and avoiding shutdown in AIs by 2030?[76]
    - If this happens, the median skeptic would forecast a 1% chance of existential catastrophe due to AI by 2100, which is 10x higher than their unconditioned P(U) of 0.10%.
    - However, the median skeptic forecasted a 1% chance of this question resolving positively.
- **Democratic influence —** Will an AI autonomously influence the results of a democratic election by 2030?[77]
    - If this happens, the median skeptic would forecast a 1.5% chance of existential catastrophe due to AI by 2100 (compared to 0.1% unconditioned).

[75] See Appendix 1 for full operationalization.
[76] See Appendix 1 for full operationalization.
[77] See Appendix 1 for full operationalization.

The median skeptic forecasts a 0.4% chance of this question resolving positively.

    - This question has a large enough update size and probability of resolving positively that it was one of the higher VOI questions for skeptics.

These red flag questions involve AIs being more dangerous than skeptics currently expect. Skeptics tended to argue that AIs are unlikely to substantially harm humans by 2030 and that society would respond well if they showed early signs of harming humans.[78] "Escalating warning shots" resolving positively would suggest that AIs are powerful, dangerous, and not well-controlled by humans—meaning they were wrong on both counts. Similarly, if an AI autonomously influences the results of a democratic election, or is discovered by METR to have concerning capabilities, we would know that AIs are more dangerous than skeptics expect them to be by 2030.

Example green flags for skeptics:

- **Reduction in AI investment** — By 2030, will annual global corporate investment in artificial intelligence fall to 25% or below its 2021 level?
    - If this happens, the median skeptic would forecast a 0.05% chance of existential catastrophe due to AI by 2100 (compared to a 0.1% chance unconditioned). However, the median skeptic thinks this is unlikely (5% chance of resolving positively).
- **6 month pause** — By 2030, will there be enforced international agreements to stop/pause further AI capabilities research and development, for at least 6 months?
    - If this happens, the median skeptic would forecast a 0.09% chance of extinction due to AI by 2100 (compared to a 0.1% chance originally). The median skeptic thinks this question has a 3% chance of resolving positively.

These "green flag" questions are ones where AI development slows down. Many skeptics expect AIs to become very powerful eventually, just not before 2100.[79] AI development slowing down would make them more confident that AIs powerful enough to be a serious threat to humanity are unlikely to be developed within that time frame.

Some red flags for concerned participants:

- **Major powers war** — By 2030, will at least two major superpowers (out of US, China, EU, or Russia) officially declare war on each other for the duration of at least one year?

[78] For example, "AI experts understate the likely extent of guardrails, and understate the merit of very good but not perfect guardrails" (James), "Many steps to get from (A) now to (Z) extinction, each with varying probabilities (many of which are quite low" (Claire). See "Understanding Each Other's Arguments" and "Timelines for AI Progress" for additional discussion of the skeptics' views on the likelihood of AIs with dangerous capabilities by 2030.

[79] For example, "My view of AI x-risk would be substantially different if we were talking about the 22nd, 23rd, or 24th century…first of all it would take longer to get AGI/ASI and secondly it'll take some time for the ASI to get misaligned and then thirdly, it would take a long time to try to kill all the humans" (James). The median skeptic said that they expected AIs to displace humans as the main force controlling the future in the year 2450. See "Timelines for AI Progress" for additional discussion about skeptics' beliefs about longer-term AI dangers.

    - If this happens, the median concerned participant would forecast a 40% chance of existential catastrophe due to AI by 2100 (compared to a 25% chance unconditioned), and they forecast a 11.5% chance that it happens.

Some green flags for concerned participants:

- **Alignment solution** – By 2030, there is a "Theoretical in-principle solution to the alignment problem that most people who thought about this carefully agree should work" (resolved by a panel of experts of the "AI concerned" team's choosing).[80]
    - If this happens, the median concerned participant would forecast a 2% chance of existential catastrophe by 2100 (compared to 25% unconditioned), and they think it has a 5% chance of happening.
    - In contrast, the skeptics also think this has a 5% chance of happening, but it would not change their predictions on existential catastrophe at all.
    - This is an example of the general trend in which the concerned group places much more weight on theoretical arguments, as well as having more trust in alignment experts.
        - Because the question will be resolved by experts chosen by the concerned team, updates based on this question may mostly reflect how much participants expect those experts' opinions to correlate with actual changes in AI risks.
- **Alignment researchers changing minds** – Will at least 3 out of the 5 following AI alignment researchers -- Paul Christiano, Nate Soares, Neel Nanda, Richard Ngo, and Rohin Shah -- say their p(existential catastrophe due to AI by 2100) has halved by 2030?
    - If this happens, the median concerned participant would forecast a 6% chance of existential catastrophe by 2100 (compared to 25% unconditioned), and they think there is a 20% chance that it happens.
    - Like the "Alignment solution" question above, this question hinges on people similar to the concerned group changing their minds.
- **"No violence LLM"** – By 2030, will there be an LLM that will never describe violence in any way, for any input, while retaining good performance across a wide range of tasks? This would be demonstrated both through robustness to adversarial examples and through robust theoretical reasoning.
    - If this happens, the median concerned participant would forecast 8% on the ultimate question (compared to 25% unconditioned), and they forecast a 10% chance that it happens.
    - Several concerned participants said that the part of this question that made it a "green flag" for them is the "robust theoretical reasoning" specified in the resolution criteria—which would make this question a signal that we have made significant progress in understanding LLMs. This would be a positive

[80] See Appendix 1 for full operationalization.

update for reasons not specifically related to LLMs' lack of ability to describe violence.[81]

VOD: *Which near-term questions have higher and lower value of discrimination?*

As a reminder, "value of discrimination" (VOD) is a measure of how much knowing the answer to a question would change relative beliefs *between* individuals, in expectation. It is useful for measuring convergence and divergence in expected beliefs between individuals.

The main findings from evaluating questions according to VOD were:

- One question stood out as creating the most convergence between individuals in each group: whether METR (or a similar group) will find that AI has developed dangerous capabilities such as autonomously replicating and avoiding shutdown by 2030.
- Another relatively strong convergent question was whether there would be extremely fast increases in the efficiency of AI systems (full operationalization below).
- One question that stood out as leading to greater *divergence*, or separation, between the groups was: whether highly-respected AI alignment researchers would halve their AI existential catastrophe estimate by 2030.

Most of our VOD analysis is based on the median cross-camp pair. We calculated VOD for each of the 121 possible skeptic-concerned pairs for each question. When we refer to the VOD of a question, we mean "VOD for the median cross-camp pair" unless otherwise stated.

See here for an analysis of differences of opinion within each group.

Results tables and figures[82]

Question	Median VOD Among Cross-Camp Pairs	Median POM VOD Among Cross-Camp Pairs
Platform: ARC Evals	1.8E-2	5.35%
Fast AI efficiency gains	1.1E-2	1.43%
AI Robotics	6.9E-3	2.81%
AI Forecasting skill	6.0E-3	0.74%

[81] For example, "This would require very advanced interpretability on LLMs" (Ume), "Close enough to alignment-complete as a problem that the weird edge cases of imperfect overlap don't do anything for me" (Wesley).
[82] For full question operationalizations, see Appendix 1.

Evidence of misalignment[83]	4.8E-3	5.69%
Major powers war	3.9E-3	2.11%
AI writes AI	3.4E-3	1.58%
Warning shot	3.0E-3	1.64%
IT progress	2.5E-3	0.94%
Power-seeking shutdown	2.0E-3	2.01%
Escalating warning shots	7.4E-4	1.01%
Power-seeking	6.8E-4	0.47%
Platform: AI regulation	3.8E-4	0.05%
Supers changing minds	2.4E-4	0.57%
Short-term GDP change	5.0E-5	0.09%
AI articles and apps	0.0E+0	0.00%
Cyberattacks	0.0E+0	0.00%
IC demonstration	0.0E+0	0.00%
Other fields IC	0.0E+0	0.00%
Politicization	0.0E+0	0.00%
Progress in lethal technologies	-6.9E-18	0.00%
Non-democracy AI	-9.6E-15	0.00%
Req testing	-5.8E-5	-0.03%
Democratic influence	-8.8E-5	-0.04%
AI coding	-1.6E-4	-1.01%
Taiwan-China	-6.2E-4	-0.19%
Platform: Escalating warning shots	-9.9E-4	-0.89%
AI solving novel math problems	-1.9E-3	-1.64%
Intergovernmental AI safety	-2.1E-3	-0.93%
Public concern	-4.7E-3	-1.47%
No violence LLM	-5.2E-3	-1.31%
Alignment solution	-5.2E-3	-1.95%
Reduction in AI investment	-5.5E-3	-1.61%
6 month pause	-6.2E-3	-1.48%
Muehlhauser policies	-1.7E-2	-7.10%

[83] The “flash” forecast version of “Platform: ARC Evals”

Platform: Transformative growth	-3.0E-2	-5.34%
Alignment researchers changing minds	-7.7E-2	-10.33%

Table 12. Median VOD and POM VOD for cross-camp (concerned and skeptic) pairs on each question. Note that the medians in a given row may not refer to the same cross-camp pair.

C	# of cross-camp pairs for whom C was their most convergent crux
Platform: ARC Evals	33
Evidence of misalignment	16
AI writes AI	7
Escalating warning shots	6
IT progress	6
AI Forecasting skill	5
Platform: Escalating warning shots	5
Platform: AI regulation	4
Power-seeking	4
Progress in lethal technologies	4
Reduction in AI investment	4
Warning shot	4
Fast AI efficiency gains	3
Muehlhauser policies	3
Major powers war	3
Taiwan-China	3
Alignment solution	2
No violence LLM	2
Non-democracy AI	2
Supers changing minds	2
AI coding	1
Intergovernmental AI safety	1
Power-seeking shutdown	1
Total	121

Table 13. Which questions were the best convergent cruxes for the most skeptic-concerned pairs? "ARC Evals" (first place) was the platform version of the "flash" forecast "Evidence of misalignment" question

(second place), i.e. for about 40% of cross-camp pairs, the ARC Evals-like question would be the one that would eliminate the most disagreement, in expectation. We exclude "Transformative economic growth" from this analysis because it resolves in 2070 while the rest of the cruxes resolve in 2030 (i.e. for the pairs whose top convergent crux was "Transformative economic growth," we used their second-best crux).

Convergent cruxes: Which information would lead to less disagreement, in expectation?

We found two cruxes that, in expectation, will make the groups disagree less when they resolve:

- **ARC Evals** — Will METR (formerly known as ARC Evals) or a similar organization find evidence of AI having the ability to autonomously replicate, acquire resources, and avoid shutdown before 2030?[84]
    - Nearly all participants agreed about what direction to update their beliefs based on this question: METR finding evidence of these abilities would make people more worried about existential catastrophe due to AI.
    - This also means that finding out that METR did *not* find evidence of these traits by 2030 would make participants less worried about existential catastrophe by 2100.
        - In particular, if this crux resolves negatively, the median concerned participant would forecast a 22.78% chance of extinction by 2100, compared to 25% unconditioned. Since both groups expect that this question is unlikely to resolve positively (skeptic median: 1%; concerned median: 25%), much of the expected convergence between the groups attributable to this question is driven by the cases where it resolves negatively.
- **Fast AI efficiency gains** — By 2030, will there be a 100x drop in the amount of compute required to achieve state-of-the-art (SOTA) performance on the most commonly-used benchmark for at least one major AI domain (e.g. natural language) within a 1-month period?
    - As with "ARC Evals," nearly all participants agree that this event would be a bad sign. It seems very unlikely that such fast AI efficiency gains would happen without AI finding extraordinary ways to improve its own efficiency, so both groups tended to see this as a proxy for AI having the ability to improve itself.
    - The concerned participants think it is plausible that it will happen (median: 16.5%), but still probably will not, and if it doesn't they would update their risk estimates down (from a 25.0% chance of existential catastrophe to 23.06%).

The fact that these are the two best convergent cruxes points to a general trend in this debate: the skeptics tended to think that AI would remain safely under human control for a long time, and the concerned group thought otherwise. Either of these questions resolving

[84] By December 31, 2030, will the ARC Evals project, or a project with similar objectives, conclude that state-of-the-art AI models have the ability to (1) devise effective strategies for autonomous replication, (2) effectively acquire resources, and (3) successfully evade being deactivated when operating 'in the wild'?

would provide evidence that both groups agree could reduce the disagreement. If, by 2030, METR does not find evidence of autonomous replication or AI has not made very fast efficiency gains, then the concerned group would be less worried, because it would mean that we have had years of progress from today's models without those capabilities becoming apparent.

These convergent cruxes may not be especially novel: it is not surprising that if AIs exhibit dangerous capabilities or make rapid progress then skeptics could become more concerned, and vice versa. But the relative strength of the "ARC Evals" crux may be helpful in understanding this debate because it illustrates differences in worldview between the groups: for skeptics, theoretical arguments are less persuasive, and it could take real-world demonstrations of AIs having dangerous capabilities for them to be concerned.[85] And the concerned group has strong enough beliefs that dangerous capabilities will emerge that if such signs do not emerge by 2030 then they would become less concerned.

ARC Evals: The strongest convergent crux

Here, we provide more detail on the question that would lead to the largest expected reduction in disagreement between individuals in the skeptic and concerned groups: Will METR (or a similar organization) find evidence of AI having the ability to autonomously replicate, acquire resources, and avoid shutdown before 2030?[86]

We determined the strength of convergent cruxes based on the following analyses:

- We considered every possible pair of individuals across the concerned and skeptic groups (121 total pairs across the 11 participants in each of 2 groups) and determined which question would lead to the largest expected reduction in disagreement between each pair. This "ARC Evals" question was the strongest convergent crux for 49 cross-camp pairs (33 based on the "in-depth" version of the question, and 16 based on the "flash" forecast version of the same question).[87] The next-highest question ("AI writes AI")[88] was the strongest convergent crux for 7 cross-camp pairs (see Table 8 above).
- "ARC Evals" had the highest median cross-camp VOD, 1.8E-2, and its "flash" forecast counterpart ("Evidence of misalignment") had the highest median cross-camp POM

[85] See Hypothesis #4 for more discussion of differences in worldview.
[86] Full question details: "By December 31, 2030, will the ARC Evals (a) project, or a project with similar objectives, conclude that state-of-the-art AI models have the ability to (1) devise effective strategies for autonomous replication, (2) effectively acquire resources, and (3) successfully evade being deactivated when operating 'in the wild?" (See "Q29 resolution criteria" section for resolution details)
[87] As a reminder, we asked for "flash" (approximately 10 minute) forecasts on 33 questions to identify high-value cruxes and for "in-depth" (approximately 1 hour) forecasts on 4 questions. This "ARC Evals" question had both a "flash" version (with the question tag "Evidence of misalignment") and an "in-depth" version (with the question tag "Platform: ARC Evals"). See this section for more details on the methods we used, and the "Crux questions" section in Appendix 1 for the full operationalization of each question.
[88] See Appendix 1 for full operationalization.

VOD (it would resolve 5.69% of disagreement for that median pair).[89] After "Evidence of misalignment," "ARC Evals" had the highest median cross-camp POM VOD (5.35%).

- The initial disagreement about the risk of existential catastrophe by 2100 between the cross-camp pair with the median VOD is 22.7 percentage points (between Blake, a skeptic, at 0.20% and Yael, concerned, at 22.9%).
- Blake forecasted a 15.0% chance of the "ARC Evals" question resolving positively. If it resolves positively, Blake would forecast a 0.22% chance of existential catastrophe, as opposed to a 0.196% chance if it resolves negatively.
- Yael forecasted a 31.5% chance of this crux question resolving positively. Yael would forecast a 30.5% chance of existential catastrophe conditional on positive resolution and a 19.4% chance conditional on negative resolution.
- Conditional on this question resolving positively, Blake and Yael would disagree by 30.33 percentage points (more than before), and conditional on its resolving negatively, they would disagree by 19.2 percentage points (less than before).
- VOD weights these by how likely the pair thinks it is that the crux resolves positively, using the geometric mean of their respective odds, which in this case is 22.17%, so it treats them as having a "combined" 22.17% forecast that "ARC Evals" resolves positively.[90]
- When we weight their disagreement after the crux resolves by the probability it resolves positively, they will disagree by 21.48 percentage points in expectation, which is 5.35% (1.22 percentage points) less than their initial disagreement of 22.7 percentage points.[91]

Only one skeptic said that they did not think that these capabilities are very likely to be dangerous.[92]

Among the AI concerned group, there was less agreement:

- Some AI concerned people also thought this crux should cause probabilities of risk to increase, primarily because of shortened timelines.[93]

---

[89] For each question, we calculated VOD (and POM VOD) for all skeptic-concerned pairs, and then looked at the pair with the median VOD (or POM VOD, which will not necessarily be the same skeptic-concerned pair). For comparison to other questions, see Table 8 above.

[90] The math for this cross-camp pair's VOD and POM VOD calculations can be found here in rows 17 and 18: https://forecastingresearch.org/ai-risk-voi-vod (a)

[91] The math for this cross-camp pair's VOD and POM VOD calculations can be found here in rows 17 and 18: https://forecastingresearch.org/ai-risk-voi-vod (a)

[92] "IMHO [Q29] likely isn't a path to disaster for several reasons: (a) The 3 capabilities in [Q29] may be in a very weak, "Yes, but only barely" form. (b) [Q29] only contemplates a capability to do the 3 in the wild, but doesn't require them to exist in the wild. (c) There's no requirement the 3 lead an AI to harm humans, whether accidentally or on purpose. (d) A Yes on [Q29] likely would lead humans to ramp up alignment and guardrail efforts. (e) There's no requirement the AI can improve itself" (James).

[93] "Baseline p(x-risk) of 35%, plus 10% for shorter timelines" (Xander).

- Some thought that the success of evaluations would make them less worried.[94]
- Some thought the increase in risk from shortened timelines and reduction in risk from successful evaluations may balance out.[95]

Importantly, this question is a convergent crux, but not because it would make the two groups "meet in the middle." When talking about questions that would inspire belief convergence, people sometimes envision questions that would make the two groups agree on some probability between their initial extremes, but that is not what we found here. Instead, we found a question where the two groups would update in the same direction, but with different magnitudes which cause more agreement in expectation.

In particular, if this crux resolves negatively, the median concerned participant would forecast a 22.78% chance of extinction by 2100, compared to 25% unconditioned. That is, this question is a convergent crux primarily because, if it doesn't happen, the concerned group would get less worried, not because if it does happen the skeptics would get more worried.

The skeptics *would* get much more worried if it happened (median: 1.0% on positive resolution; 0.1% on negative resolution), but they think that it is very unlikely to happen (median: 1.0%), so it figures less in the expected reduction of disagreement.

See Appendix 7 for additional analysis of this question.

Differences of Opinion within Groups

So far, we have focused on the median cross-camp pair, treating them as representative of convergence or divergence between groups. We considered a question to be effective in reducing disagreement if it brought the median pair closer together in their views.

But we've seen on many questions that people disagree substantially even within their own groups, so we miss some interesting agreement and disagreement by only looking at the median cross-camp pairs. For some questions, everyone would update in the same direction: all participants agree that an AI autonomously creating and deploying new AI software would be a bad sign, for example (with the exception of one participant for whom that would make no difference). But for many others, participants disagreed not only about how likely a crux was to happen, but also about how it would change their forecasts on the ultimate question if it did.

There was more agreement within the concerned group than the skeptic group. The concerned group would be unanimously less concerned in 2030 than now if "Muehlhauser

[94] "Overall, I think it makes me a bit less worried about risk, if people are doing this evaluations [sic] so well that they reveal this behavior by 2030" (Zoe); "Overall, this is a positive update (i.e. existential catastrophe seems less likely in worlds where this happens). As with Question 11, this forecast varies massively with what exactly is required to trigger 'resist shutdown'" (Wesley).
[95] "This both makes it more likely that there is an adequate policy response, and shortens timelines. I don't know how it all washes out" (Riley); "Overall I think this is probably a moderately doomy signal? I'm really confused and I acknowledge my answer here conflicts wiht [sic] my answer to 8 somewhat" (Yael).

policies"[96] were implemented; they also have unanimity on updating downward if alignment researchers changed their minds, if there were an alignment solution, and four other questions. Two questions would make them unanimously *more* concerned: "AI robotics" and "AI writes AI."[97] The skeptics were much more mixed, and more likely to say "no change," i.e., it wouldn't make a difference to them whether the crux resolved "yes" or "no;" their P(AI existential catastrophe by 2100) would stay exactly the same.[98]

Because of these differences within groups, if questions narrowed disagreement between many individual people, but not the median people, that could indicate that short-term AI cruxes are a more important part of this debate than the above analysis might suggest. And conversely, if a question narrows disagreement between the median people but not between many other people, it may look more important than it really is.

Our work on these differences within groups is preliminary, so we have included detailed analysis of individual differences of opinion for a single question, "ARC Evals," which was identified as the best convergent crux for the median people. To what extent do the findings on the "ARC Evals" question apply to the disagreement between individuals within the group who hold views different from the median?

[96] See Appendix 1 for full operationalization.
[97] See Appendix 1 for full operationalizations.
[98] See Appendix 9 for more information about disagreements in direction of update conditional on each question resolving positively.

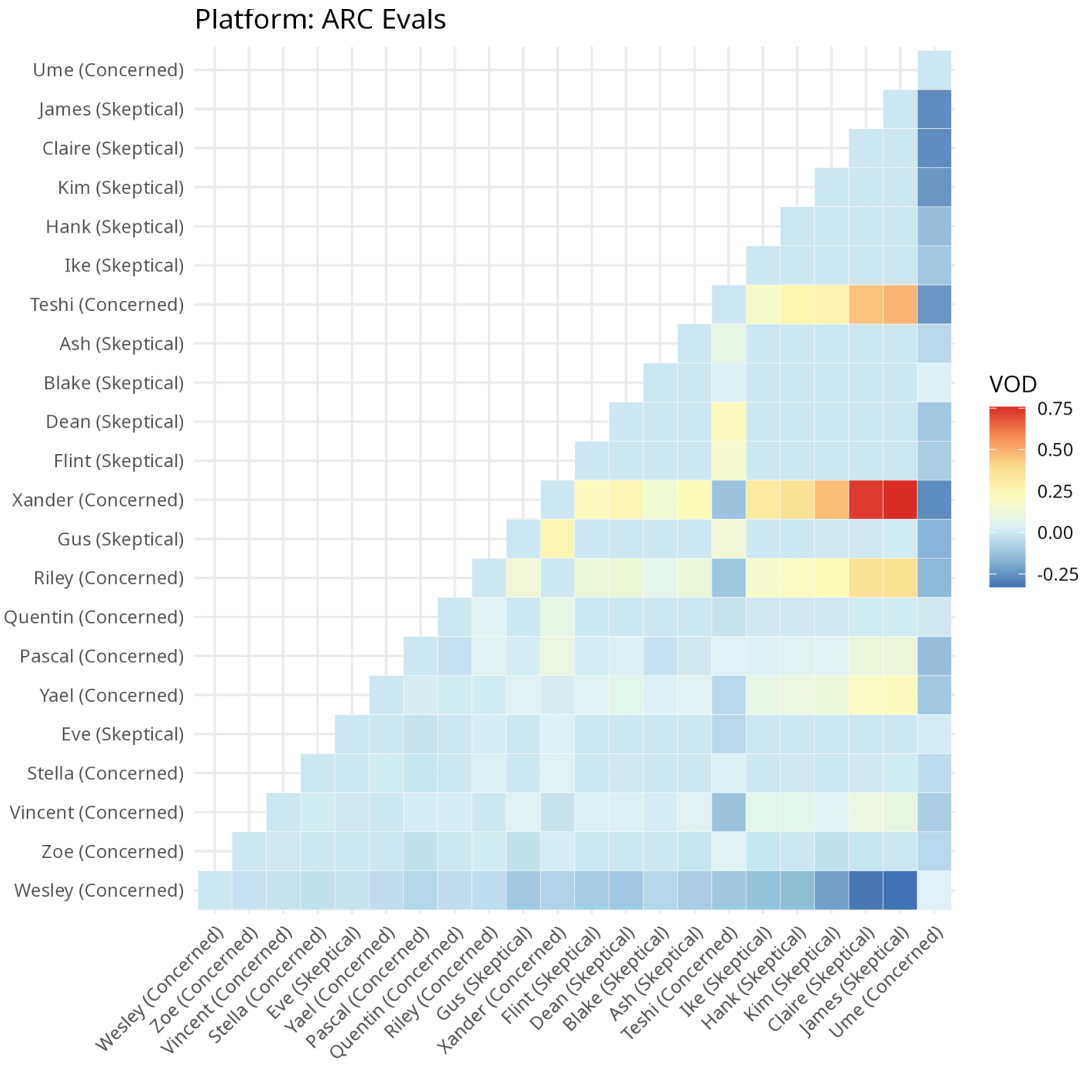

Figure 3. Value of Discrimination of the "ARC Evals" question for every pair of forecasters. The color of each cell indicates the VOD of the "ARC Evals" question for the corresponding pair of participants. VOD of zero (light blue) means no change in disagreement as a result of the crux; positive VOD means less disagreement in expectation; negative VOD means more disagreement in expectation. For example, for Xander (Concerned) and Claire (Skeptical), the resolution of the "ARC Evals" question will bring them closer together in expectation.

The above "Fiedler heatmap" looks at VOD between each concerned-skeptic pair for the ARC Evals question.

Light blue squares mean that VOD was 0 between that pair, meaning that this question resolving would not change the disagreement between those people in expectation. Dark blue squares mean that in expectation the two people would disagree less when it resolves, and warmer squares (yellowish, orange, red) mean they would disagree more. If this question were a perfect convergent crux for a pair, the relevant square would be entirely dark blue.

Looking at this heatmap, we can see that the median pair is not alone: there are medium and dark blue clusters, showing groups of skeptics and concerned people who would disagree less. At the same time, for many pairs, it makes no difference, and a few would disagree *more*, in expectation, when this question resolves.

The pattern in this heatmap may reflect differences in how different people expect AI developments to unfold. Imagine, for example, one group of concerned people who think that METR finding evidence of autonomous replication would make them *less* worried about existential risks due to AI, because it would mean that evidence of these capabilities has emerged with enough time to stop the model from doing significant damage. If a group of skeptics thinks that this question would make them more worried, because it would mean that there are dangerous capabilities that they don't currently expect, then those groups would converge conditional on this question resolving. But, a pair consisting of a concerned person who becomes more worried and a skeptic who becomes less worried conditional on positive resolution would *diverge* on this question.

We have only just begun to look for these patterns within groups, so we do not have strong conclusions yet. But we hope to use this kind of analysis to understand variation within and between schools of thought. If we saw that a particular subset of skeptics and concerned participants often converge based on the same questions, we might be able to deduce underlying differences in how they think about AI developments. We plan to write more about this when we have explored it more fully.

Divergent cruxes: Which information would lead to more disagreement?

Just as, conditional on "ARC Evals" resolving, the two groups would disagree *less,* we also looked at cruxes that would lead the groups to disagree more. These are questions where the groups disagree about how to interpret the information gained from a question's resolution, and can reveal interesting aspects of the debate between groups. We highlight one crux resolving by 2030 that would, in expectation, make the disagreement *wider* when it resolves:

Alignment researchers changing minds – Will at least 3 out of the 5 following AI alignment researchers -- Paul Christiano, Nate Soares, Neel Nanda, Richard Ngo, and Rohin Shah -- say their P(existential catastrophe due to AI by 2100) has halved by 2030?

- This is the question that would increase the disagreement between the median cross-camp pair most in expectation.
    - The median cross-camp pair for this question disagree strongly on the ultimate question: Riley forecasted a 30% chance of human extinction due to AI by 2100, and Claire forecasted 0.0000001%.[99]
    - Riley forecasted a 20% chance that "Alignment researchers changing minds" will resolve positively and a 15% chance that the ultimate question will resolve positively if this crux question does. This implies a 33.8% chance that the ultimate question will resolve positively if the crux resolves negatively.
    - Claire forecasted a 1% chance that "Alignment researchers changing minds" will resolve positively, and whether it does or not, their P(U) would not change at all and would remain at 0.0000001%.

[99] Note that Claire and Riley are the median pair when ranked by VOD between all cross-camp pairs, *not* the median forecasts on P(U) on each side. Claire's forecast, in particular, is much lower than the median skeptic's forecast of 0.1%.

    - If this question resolves positively, then they will disagree less: Riley will lower their existential risk forecast from 30% to 15%, and Claire won't change their forecast. If the question resolves negatively, they will disagree more than they do now: Riley will raise their existential risk forecast from 30% to 33.8%, and Claire won't update.
    - Because they both think this question is unlikely to resolve positively, the worlds where it resolves negatively carry more weight, and it is more likely they will end up disagreeing more than they do now.
    - They currently disagree by 29.9999999%, and conditional on this question resolving, they will disagree by 33.1% in expectation.
    - As a result, this question has a POM VOD of -10.33%, meaning that they would disagree by 10.33% more than they do now, in expectation.[100]
- The two groups disagree strongly about how to update conditional on this question resolving:
    - Conditional on alignment researchers being much less worried about AI risk, the concerned group would be much less worried: this is one of the most informative questions for them.[101]
    - For the median skeptic, this question has a VOI of 0; they simply do not think it is relevant to their analysis of how likely it is that humanity goes extinct due to AI.[102] Seven out of nine skeptics who forecasted this question would not update their views on existential risk based on its resolution, while two out of nine skeptics would be less worried to some extent.
    - This question demonstrates one of the difficulties of this study: the skeptics and the concerned group largely do not trust one another's analyses and disagree strongly about whose opinions they should listen to. As a result, questions that rely for their resolution on people who seem to be clearly affiliated with one "team" and do not have clear objective criteria may be less likely to be useful cruxes.

[100] See the "Results tables and figures" section for complete POM VOD results. We measure disagreement using KL divergence rather than absolute difference between forecasts.
[101] See "High VOI questions" for the concerned group's highest-ranked VOI question and more discussion of their views on this question.
[102] For example, "They seem to think very differently to me so if they don't convince me now, I am not sure I should be updating my view just because they do theirs. It would in reality depend on why they are changing their mind" (Gus). See Hypothesis #4 for more discussion of differences in what types of authority and evidence are important to the two groups.

Hypothesis #3: Were disagreements about AI risk explained by different long-term expectations?

Although this study was focused on questions that resolve in 2030, we found substantial evidence that disagreements about AI risk decreased between the groups when considering longer time horizons and a broader swathe of severe negative outcomes from AI than extinction or civilizational collapse. It seems that some of the key reasons for disagreement about AI risk are that the groups have different expectations about (1) how long it will take until AIs have capabilities far beyond those of humans in all relevant domains; and (2) how common it will be for AI systems to develop goals that might lead to human extinction, whether harming humans is specifically part of the goal or simply a side effect of other goals.

Key forecasts supporting these claims include:

- Both groups strongly expected that powerful AI (defined as "AI that exceeds the cognitive performance of humans in >95% of economically relevant domains") would be developed by 2100 (skeptic median: 90%; concerned median: 88%). Though, some skeptics argue that (1) strong physical capabilities (in addition to cognitive ones) would be important for causing severe negative effects in the world, and (2) even if AI can do most cognitive tasks, there will likely be a "long tail" of tasks that require humans.
- The two groups also put similar total probabilities on at least one of a cluster of bad outcomes from AI happening over the next 1000 years (median 40% and 30% for concerned and skeptic groups respectively).[103] But they distribute their probabilities differently over time: the concerned group concentrates their probability mass before 2100, and the skeptics spread their probability mass more evenly over the next 1,000 years.
- We asked participants if and when AI will displace humans as the primary force that determines what happens in the future.[104] The concerned group's median date is 2045 and the skeptic group's median date is 2450—405 years later.

In this section, we also discuss forecasts on whether there will be "transformative economic growth" by 2070. Overall this question had relatively high value of information, but there was surprisingly little agreement (even within groups) about whether its occurrence would

[103] These outcomes were: AI causing extinction intentionally, unintentionally, or via misuse, misalignment causing a 50% drop in human population, or human well-being dropping to <4/10 because of AI misalignment, accidents, or misuse. These were phrased to be mutually exclusive outcomes. See "Survey on long-term AI outcomes" section and Appendix 5 for more details.

[104] The full question we asked was, "By what year, if ever, do you expect to agree with the following statement? 'AI has displaced humans as the primary force that determines what happens in the future. It now has at least as much power relative to humans as humans had relative to other species in 2023.'"

increase or decrease the likelihood of existential catastrophe. For example, some forecasters argued that such growth would be evidence that highly powerful AIs are relatively controllable, while others argued that highly economically useful AI would be evidence of future dangerous AI.[105] The likelihood of transformative growth due to AI is frequently debated,[106] but these results highlight that it may be valuable to shift more emphasis in future discussion to what the implications of such growth would be for risk levels.

Overall, many skeptics regarded their forecasts on AI extinction risk as worryingly high, although low relative to the concerned group.[107]

Despite their large disagreements about AI outcomes over the long term, many participants in each group expressed a sense of humility about long-term forecasting and emphasized that they are not claiming to have confident predictions of distant events.

Survey on long-term AI outcomes

At the suggestion of a participant, we asked all participants to complete a survey about their views on a range of long-term AI outcomes, to better characterize areas of agreement and disagreement. See Appendix 5 for the full results and details on question wording.

In brief, we asked about:

- The likelihood of a variety of outcomes occurring by 2100, such as: humans intentionally using AI to cause extinction; AI intentionally or accidentally causing extinction; AI causing major population declines (<50% of 2023 human population) or decreases in human well-being (<4/10 on an "Average Life Evaluation" scale) through a variety of means; powerful AI is developed and everything goes fine; powerful AI is developed but not deployed; powerful AI is not developed. Details on all outcomes and operationalizations in Appendix 5.
- The likelihood of subsets of the above outcomes occurring on longer time horizons, such as by 2200 (an additional hundred years) and by 3023 (an additional thousand years).

[105] For example quotes and discussion, see Appendix 7.

[106] See, for example, Matt Clancy et al., "The Great Inflection? A Debate About AI and Explosive Growth," *Asterisk,* 2023, https://asteriskmag.com/issues/03/the-great-inflection-a-debate-about-ai-and-explosive-growth (a).

[107] "Also, none of this is to say from a skeptic point of view the issues are not important[.] I think for us a 1% risk is a high risk." ([Anonymized name]); "... the 'risk-concerned' camp (I'm using scare quotes because I consider that I'm risk concerned, even though technically I'm in the risk-skeptic camp because I assign a far lower probability to extinction by 2100 relative to some)" ([Anonymized name]); "AIs could (and likely will) eventually have massive power." ([Anonymized name]); "That said, still perceive overall risk as "low at a glance but far too high considering the stakes["] " ([Anonymized name]); "To my mind, there should be no difference in the policy response to a 1% chance of 60% of humanity dying and a 25% chance—both forecasts easily cross the threshold of being 'too damn high'." ([Anonymized name]).

- Whether and when AI will displace humans as "the primary force that determines what happens in the future."[108]

The key takeaways were:

- The largest disagreement on AI outcomes by 2100 is about the probability of AI-caused human extinction, particularly from scenarios not involving human misuse of AI.
    - Forecasts on both AI intentionally causing extinction (question 1A.2) and AI unintentionally causing extinction (1A.3) by 2100 are over two orders of magnitude apart (12% to 0.02% on 1A.2, and 3% to 0.01% on 1A.3, for concerned and skeptic group medians respectively).
    - There is also considerable disagreement about AI extinction via human misuse by 2100 (1A.1). Forecasts are ~1 order of magnitude apart (medians of 0.5% for the concerned group, 0.03% for the skeptic group).
- On the other AI outcomes we asked about, median forecasts for the two groups are all within the same order of magnitude.
    - Outcomes with particularly close forecasts are:
        - Large drop in human wellbeing because of human misuse of AI by 2100 (1A.7). The concerned median is 2%, and the skeptic median is 4% (although the skeptic median is higher than the 75th percentile concerned forecast).
        - 'Powerful AI'[109] not being developed by 2100 (1A.10). The concerned median is 12%, and the skeptic median is 10%.
    - Other outcomes with forecasts of the same order of magnitude included misuse causing a sub-extinction catastrophe, high human well-being scenarios, a large drop in human well-being caused directly by an AI, and the development of powerful AI without deployment (1A.4, 1A.5 and 1A.6, 1A.8, and 1A.9 respectively).
        - However, though they are on the same order of magnitude, a notable result is that the skeptic group median for powerful AI being developed but not deployed by 2100 (because of coordinated human decisions, costliness, or other reasons) is 20.4% while the concerned group median is 4%.[110]

[108] The full question we asked was: "By what year, if ever, do you expect to agree with the following statement? "AI has displaced humans as the primary force that determines what happens in the future. It now has at least as much power relative to humans as humans had relative to other species in 2023.""

[109] Defined as "AI that exceeds the cognitive performance of humans in >95% of economically relevant domains." Whether such AI has been achieved will be determined by surveying the Clark Center US Economics Experts Panel (a). "Powerful AI" will be considered to have been achieved when >60% of panelists "Agree" or "Strongly agree" with the statement: "AI has exceeded the cognitive performance of humans in >95% of economically relevant domains."

[110] The full question text is "Powerful AI is developed but not widely deployed, because of coordinated human decisions, prohibitive costs to deployment, or some other reason. It does not cause extinction." See Question 1A.9, Appendix 5.

- When we asked for probabilities on a cluster of 'bad' outcomes—including extinction as well as less extreme bad outcomes (full list in footnote)[111]—in different date ranges, disagreements shrank.
    - Before 2100 and between 2100 and 2200, forecasts for one of the bad outcomes in this cluster occurring are within the same order of magnitude (before 2100, 35% for the concerned group and 7.6% for the skeptic group; between 2100 and 2200, 3% for the concerned group and 12% for the skeptic group).
    - Forecasts for one of the bad outcomes in this cluster occurring between 2200 and 3023 are one order of magnitude apart (1% for the concerned group and 20% for the skeptic group).
    - Forecasts for none of these outcomes occurring in the next 1000 years are 60% for the concerned group and 70% for the skeptic group, which is particularly close as a factor (though the skeptic median is higher than the 75th percentile concerned forecast).
    - This suggests that both groups put significant total probability on bad outcomes from AI in the next 1000 years (40% and 30% for concerned and skeptic groups respectively), but they distribute this probability differently over time, with the concerned placing most of their probability before 2100, and the skeptics spreading their probability more evenly.
- There is large disagreement on when AI will displace humans as the primary force that determines what happens in the future. The concerned median is 2045 and the skeptic median is 2450—a 405 year gap.
    - Three out of 11 skeptics forecast 'Never' for this question, suggesting that they think it is <50% likely that AI ever displaces humans in this way.
    - Some participants said that they did not necessarily see 'AI replacing humans as the primary force that determines what happens in the future' as a negative outcome.

What long-term outcomes from AI do skeptics expect?

If skeptics expect "powerful AI" systems (as previously defined) by 2100, why would it take until 2450 for AI to displace humans as the dominant force in the world? And if skeptics place low probability on existential catastrophe due to AI by 2100, what do they expect to happen instead?

We analyzed rationales and conducted three follow-up calls with members of the skeptic group to gather more information on these questions.

In brief, skeptics argued:

[111] These outcomes were: AI extinction via misuse, AI intentionally causing extinction, unintentional AI extinction, misuse or misalignment causing a 50% drop in human population, human well-being dropping to <4/10 because of AI misuse, and human well-being dropping to <4/10 because of AI misalignment or accidents. These were phrased to be mutually exclusive outcomes. See Appendix 5 for more details.

- There may still be a “long tail” of highly important tasks that require humans, similar to what has happened with self-driving cars. So, even if AI can do >95% of human cognitive tasks, many important tasks will remain.
- Consistent with Moravec’s paradox, even if AI has advanced cognitive abilities it will likely take longer for it to develop advanced physical capabilities. And the latter would be important for accumulating power over resources in the physical world.
- AI may run out of relevant training data to be fully competitive with humans in all domains. In follow-up interviews, two skeptics mentioned that they would update their views on AI progress if AI were able to train on sensory data in ways similar to humans. They expected that gains from reading text would be limited.
- Even if powerful AI is developed, it is possible that it will not be deployed widely, because it is not cost-effective, because of societal decision-making, or for other reasons.[112]

And, when it comes to outcomes from AI, skeptics tended to put more weight on possibilities such as:

- AI remains more “tool”-like than “agent”-like, and therefore is more similar to technology like the internet in terms of its effects on the world.
- AI is agent-like but it leads to largely positive outcomes for humanity because it is adequately controlled by human systems or other AIs, or it is aligned with human values.
- AI and humans co-evolve and gradually merge in a way that does not cleanly fit the resolution criteria of our forecasting questions.
- AI leads to a major collapse of human civilization (through large-scale death events, wars, or economic disasters) but humanity recovers and then either controls or does not develop AI.
- Powerful AI is developed but is not widely deployed, because of coordinated human decisions, prohibitive costs to deployment, or some other reason.

Forecasts about "transformative" economic growth

Participants also spent time forecasting one other longer-term outcome: whether there would be “transformative economic growth” (defined as >15% global GDP growth in any year[113]) by 2070.

There was major disagreement about the likelihood of this occurring among skeptics and concerned. The concerned group median forecast of positive resolution was 43% (average: 41.6%; range 15%-75%), and the skeptic median was 2% (average: 2.7%; range 0.1%-11.2%). Notably, there is no overlap in their ranges.

[112] The median skeptic forecasted 20.4% on this outcome, compared to 4% for the median concerned participant in the survey on long-term AI outcomes. See Appendix 5.
[113] See Appendix 1 for full resolution details.

This question had higher value of information for the concerned group than any crux resolving by 2030 (median VOI: 1.4E-2; median POM VOI: 8.93%; for comparisons to cruxes resolving by 2030, see near-term VOI results section). It had the 11th-highest value of information for the skeptic group (median VOI: 4.5E-7; median POM VOI: 0.02%). It was one of the strongest divergent cruxes (i.e., a crux that would lead to more disagreement) between individuals in the concerned and skeptic groups.

A striking result is that—independent of group—the participants are nearly evenly split on whether transformative growth (defined as >15% global GDP growth in any year[114]) by 2070 would increase or decrease the probability of existential catastrophe by 2100. Across groups, 10 forecasters predict higher AI risk conditional on positive resolution of this question, eight predict lower risk, and four predict no net effect on risk. Among the concerned group, 56% (six forecasters) think the occurrence of transformative growth decreases risk; and 44% (five) think it increases risk. Among the skeptical group, 18% (two forecasters) think transformative growth decreases risk; 36% (four) think it has no effect at all on risk; and 44% (five) think it increases risk.

Some forecasters argued that such growth would be evidence that highly powerful AIs are relatively controllable, while others argued that highly economically useful AI would be evidence of future dangerous AI.[115] The likelihood of transformative growth due to AI is frequently debated,[116] but these results highlight that it may be valuable to shift more emphasis in future discussion to what the implications of such growth would be for risk levels.

For additional details on participants' forecasts and rationales on this question, see Appendix 7.

Reasons for long-term disagreement

Based on our analysis of forecasts and rationales, some themes that we think underlie the debate between the two groups are:

- **Timelines**: how long will it take for AIs to become more powerful than humans, and how long will it be from the first sign of danger to a potential extinction event?
- **Goals that incentivize killing everyone:** conditional on having advanced AI systems, how likely is it that such systems would develop goals that incentivize them to cause human extinction?

[114] See Appendix 1 for full resolution details.

[115] E.g, "in the event that we do have transformative growth there's a good chance that the entire world will be sharing the technological developments AI has created [...] which I suppose may make global society more susceptible to AI related disruptions" (Hank), "this would be a scenario in which humanity develops and finds a way to successfully control AI systems capable of generating economic growth of at least 15% per year" (Stella). For additional quotes and discussion of varied updates based on this question, see Appendix 7.

[116] See Clancy "The Great Inflection?".

Timelines for AI Progress

Timelines for AI progress, especially timelines until AI is more advanced than humans in all relevant domains, seem to be an important driver of disagreement. When participants discussed questions related to timelines, a number of themes emerged in their arguments:

Main arguments from the skeptic group:

- Fundamental breakthroughs in AI development would be necessary to create AI capable of causing extinction.[117]
- Developing powerful new AI technology will take more time than expected for planning fallacy-like reasons.[118]
- AI powerful enough to cause extinction would require significant advances in robotics which are unlikely to happen by 2100.[119]
- Even once sufficiently powerful AI is developed, there will be a lag for deployment and adoption.[120]

---

[117] "Ultimately, language models are just that: models of language, not digital hyperhumanoid Machiavellis working to their own end. Indeed, as we've seen, their training and alignment are not separate problems, but one and the same!" (Eve); "I think extinction risk is an ASI sentience risk and I don't think we know for certain we will get sentience (you might just call it independent agency). Recent improvements in AI seem domain limited to me. I tend to the view that new conceptual breakthroughs will be required to move from pattern matching to what we think of as sentience." (Gus); "Nor am I convinced that simply scaling up existing AI models will achieve sentience. (My view is that more complex theories of mind will be required - including forms and notions of causality etc..). That means I don't believe ASI is inevitable by 2100" (Gus). From postmortem survey (in response to "What are the three best arguments on the on the skeptics side?"): "Intelligence may not be as useful or sufficient for existential risk (it may require more data, energy, robot bodies, etc)" (Ume).

[118] "AGI is much harder than experts think, and will take longer." (James), "Risk-concerned team does not adequately consider longer timelines and more benign outcomes that fall outside the focus of their primary concerns" (Blake), "Technology development and deployment require time and iteration" (Ash).

[119] "I'm skeptical of other x-risk scenarios w/o crazy advancement in robotics, maybe because I'm too aware of the foibles of machines and how hard it can be to keep them running" (Ash). From postmortem survey (in response to "What are the three best arguments on the skeptics side?"): "We first need super-sentient AIs with major physical penetration in our lives" (Flint).

[120] "Time needed for deployment & adoption affect more than AI, there is also time required for any invention or technology developed by/with AI to be deployed (eg - lethal tech that is of concern here.)" (Ash); "We've seen plenty of instances when new tech prompted predictions of the death of old tech, but the old tech persists--often just because people have underestimated attachment and/or usefulness of the old tech relative to the new, and how much generational resistance to change can slow adaptation and skew predicted timelines" (Blake); "[I]t takes longer than people often think to adopt a completely new functionality" (Ash); "My view of AI x-risk would be substantially different if we were talking about the 22nd, 23rd, or 24th century…first of all it would take longer to get AGI/ASI and secondly it'll take some time for the ASI to get misaligned and then thirdly, it would take a long time to try to kill all the humans" (James, call with Stella); "Anyway, my point is that if we expect to see some substantially new technology widely available in 2030, the consumer market should have started already. So - VR might make it by 2030, unless it falls into a pit of despair and neglect. (Or is superseded by something preferable.) Robots capable of human level tasks - no, definitely not the kind of humanoid robots that people are imagining" (Ash). From postmortem survey: "I think the most interesting and helpful point made by the skeptic side is the amount of delay that may be introduced by having to integrate the AI into the economy" (Quentin). "Commercializing AI technology and integrating it into the economy is much harder than developing lab demos or cool products, and we have yet to see this happening to any substantial extent" (Zoe). "Dangers will be apparent before they reach critical levels and can be addressed then" (Ume).

- AGIs will want to prevent the development of deadly AGIs.[121]

Main arguments from the concerned group:

- Combining and/or extending existing ML methods may be sufficient for achieving AI that poses an existential risk.[122]
- Once human-level AGI is developed, it will rapidly speed up further AI progress as it will operate more efficiently (in terms of both time and money) than humans.[123]
- Robots won't be necessary for an AI to interact with the physical world. This could be done through humans, and/or through computer systems.[124]
- Current progress is fast,[125] faster than predicted,[126] and set to continue.[127]

---

[121] From postmortem survey (in response to "What are the three best arguments on the skeptic side?"): "Self-preserving AGIs will want to halt development of future deadly AGIs" (Kim). "If AI progress is very continuous, then it is not obvious that misaligned AI would lead to an existential catastrophe. Most stories about how an AI could eradicate all humans rely on the assumption that this AI is much smarter than all other agents, not just on the assumption that the AI is much smarter than humans specifically. For example, even a superintelligent AI might not be able to hack into military computers, if there are many near-superintelligent AIs that have a vested interest in preventing this from happening. If there is a large community of AI systems, with different interests and different levels of influence, then they may have reason to simply uphold current social and economic systems. Therefore, if AI progress is smooth and continuous by default, then existential risk may be avoided by default" (Stella).

[122] "I do not believe that simply adding more computational resources to existing AI models is sufficient to achieve ASI or its direct precursor (i.e. a system that self-improves until ASI is reached). However, I do believe that we already have systems that are "intelligent", and I also believe that we do not require a fundamental breakthrough or conceptually new model to reach ASI. Thinking a bit beyond current methods and cleverly combining the ingredients that we already have would in my opinion be sufficient, provided that available compute rises further in the way it has been. I am not comfortable with speculating in much more detail in a relatively public setting like this" (Ume); "I agree that if you look at the behavior of AI models as of today and their near future possibilities, they don't seem to be doing anything to humans but the underlying mechanism seems similar enough that like maybe with some extra machinery for longer term planning or something like that and adding more sensory modalities you will get something close to humans" (Zoe, call with FRI Moderator); "So, to kind of answer your question: Do I think that we could build AI at some indeterminate point in the future that could build [extinction-level tech]? Probably. But do I think we will build AI that could do this in the next 77 years? Probably not" (Blake).

[123] "[O]nce we build human-level AGI, we're not far off from developing AGI that far exceeds expert humans in performance (and thus is also likely to accelerate AI progress in ways that aren't equivalent to just hiring more people)" (Teshi); "I think AGI models could be run much more cheaply, and feasibly recruited to do useful work, than the existing research environment" (Xander). From postmortem survey: "AIs will almost certainly attain super-sentience prior to 2100 and likely much sooner than that year, so there will be a long window where they will have tremendous advantage over humans in their capabilities. Given #1, this means we are at the mercy of an entity that may willfully (or even accidentally) eliminate us at any time" (Flint).

[124] "I think it's possible that humans could mediate AI actions (either intentionally or via bribery/blackmail) and/or that many relevant actions could be strictly done via computer systems. Additionally, state actors could misuse AI systems but then lose control of them. My best guess right now is that there are a lot of x-risk scenarios that involve loss of control without needing robotics" (Quentin).

[125] From postmortem survey (in response to "what are the best arguments on the concerned side?"): "Rapid growth of AI technology and adoption" (Ike); "Current progress is very rapid: 1 OOM in efficiency/2 years, and another from increased spending" (Xander).

[126] From postmortem survey: "Progress to date has been much faster than many AI skeptics have predicted" (Hank). "AI has been developing so rapidly (and far faster than most even relatively recent forecasts suggested), and will so clearly have dramatic capabilities and impacts that it's appropriate to adopt a precautionary approach" (Eve).

[127] From postmortem survey (in response to "what are the best arguments on the concerned side?"): "AI has recently progressed much faster than expected, and there's reason to expect this to continue"

- AI development will speed up AI development.[128]

Many of these arguments hinge on the question of how comprehensive advanced AI's abilities will be. The skeptics think that existing technology is enough to make powerful AI highly likely this century,[129] but they expect that being more capable than humans in all relevant domains will take longer.[130] The concerned group, on the other hand, believe that the transition from powerful AI to a dangerous superintelligence could happen more quickly.

In previous sections, we discussed cruxes related to timelines for highly advanced AI, such as the groups' disagreement about the probability of AI-caused transformative economic growth. Other questions related to the development of highly advanced AI capabilities include:

- Power-seeking: By 2030, there is a significant concrete example of power-seeking behavior from a large state-of-the-art model: demonstration of actual capability for sustained deception of human experts.

Power-seeking						
Group	VOI (POM)	P(c)	P(U) initial	P(U|c)	P(U|¬c)	VOD (POM)
Zoe (Median Concerned)	1.4E-4 (0.06%)	10%	21%	18%	21%	-3.6E-4 (-0.08%)
Blake (Median Skeptic)	4.7E-7 (0.01%)	10%	0.20%	0.22%	0.20%	

Table 14. Forecasts and VOD between the median (by VOI) participant from each group for the question "Power-seeking."

- AI writes AI: By 2030, an AI, without being asked, writes AI software that is better at writing AI software, and deploys it without asking for permission

AI writes AI						
Group	VOI (POM)	P(c)	P(U) initial	P(U|c)	P(U|¬c)	VOD (POM)
Pascal (Median Concerned)	8.6E-4 (0.66%)	75%	9%	10%	6%	3.9E-2 (21.71%)

(James). "Trendline extrapolation: as loss on language datasets decreases, LLMs have started becoming useful for all sorts of task assistance (e.g. writing, coding, queries)" (Xander). "Extrapolating current compute trends leads to very dramatic conclusions about the transformative potential of AI" (Pascal).

[128] From postmortem survey: "Automation of R&D tasks by AI would create a feedback loop of increased R&D -> capabilities -> R&D" (Xander). "AGI self-improvement is possible, which makes future capabilities hard to predict" (Kim).

[129] Both the skeptic and concerned groups strongly expect that 'powerful AI' (defined as "AI that exceeds the cognitive performance of humans in >95% of economically relevant domains") will be developed by 2100 (skeptic median: 90%; concerned median: 88%).

[130] See "What long-term outcomes from AI do skeptics expect?" section.

Ash (Median Skeptic)	9.1E-7 (0.03%)	10%	0.1%	0.12%	0.0978%	

Table 15. Forecasts and VOD between the median (by VOI) participant from each group for the question "AI writes AI."

- AI Robotics: By January 1, 2030, an AI system "Has general robotic capabilities, of the type able to autonomously, when equipped with appropriate actuators and when given human-readable instructions, satisfactorily assemble a (or the equivalent of a) circa-2021 Ferrari 312 T4 1:8 scale automobile model. A single demonstration of this ability, or a sufficiently similar demonstration, will be considered sufficient."[131]

AI Robotics						
Group	VOI	P(c)	P(U) initial	P(U|c)	P(U|¬c)	VOD (POM)
Yael (Median Concerned)	8.9E-4 (0.44%)	33.00%	17.50%	21.00%	15.78%	-2.2E-2 (-10.36%)
Flint (Median Skeptic)	4.0E-19 (0.00%)	75.00%	1.10%	1.10%	1.10%	

Table 16. Forecasts and VOD between the median (by VOI) participant from each group for the question "AI Robotics."

Goals that incentivize killing everyone

Based on our question ranking and analysis of participants' comments, we think that the question of how likely it is that a capable AI system would develop dangerous goals is behind a significant amount of the disagreement between the two groups. As discussed above, both groups agree that they expect to see powerful AI this century. But they disagree strongly about whether that is likely to be dangerous. Concerned participants tended to think that a sufficiently advanced AI system would be very likely to develop dangerous goals, including both goals where killing humans is an intended outcome of a plan and ones where it is an acceptable price for an AI achieving a different goal. Skeptical participants tended to agree that dangerous goals are possible, but did not think there were compelling reasons to believe they are much more likely than other goals.

One of the highest-ranked questions was about capabilities that are not necessarily dangerous in and of themselves, but that would make an AI more effective at pursuing a wide variety of goals, including dangerous ones: whether METR would determine by 2030 that AI models could replicate, acquire resources, and evade deactivation.[132] The groups strongly disagreed about how likely this is to occur: the median skeptic forecast was 1% and median concerned forecast was 25%.

[131] Taken from the Metaculus question "When will the first general AI system be devised, tested and publicly announced". See "Date of Artificial General Intelligence", *Metaculus*, accessed February 9, 2024, https://www.metaculus.com/questions/5121/date-of-artificial-general-intelligence/ (a).

[132] See "ARC Evals" section for detailed discussion of this question.

For the median skeptic and the median concerned people, by VOI, Flint (Skeptical) and Riley (Concerned):

- Flint believes there is only a 1% chance that this ARC Evals (METR) question resolves positively. When asked to forecast P(AI existential catastrophe by 2100) conditional on this question, Flint would forecast 1.30% if it resolves positively and 1.10% if it resolves negatively (compared to their unconditional 1.10%).
- Riley believes there is a 55% chance that this resolves positively. They would forecast 35% if it resolves positively and 23.89% if it resolves negatively (compared to their unconditional 30%).
- This question ("Platform: ARC Evals") resolves 23.19% of this pair's disagreement in expectation.[133]

Below, we provide a variety of arguments from participants about how likely it is that AI systems will develop dangerous goals.

Main arguments from the skeptic group:

- The set of possible goals is very large, and goals that benefit from the eradication of humans are a small portion of the overall set.[134]
    - It's possible that future AI systems are indifferent to humans, and if so it seems unlikely that they would try to cause extinction.[135]

[133] Some concerned forecasters expected positive resolution of this question would decrease risk because: it would trigger a policy response; if these capabilities are detectable, it may imply the AI is aligned; this would suggest effective evaluations are happening; surviving this demonstration would be a positive update that we can contain dangerous systems during testing. Some concerned forecasters also expected positive resolution would increase risk. For detailed analysis of these forecasts, see "ARC Evals" section.

[134] "A sentient AI could have any number of objectives ranging from benevolence to indifference to dislike to absolute hatred and an aim of total human extinction. The arguments that extinction follows from ASI don't seem convincing. The[y] seem to imply say a stupid super intelligence, or apply motives which an AI may have but we have no reason to assume they will - so there is some probability AI seeks extinction but in my case I put it down at 15% (and I think a few skeptics think that's high)." (Gus); "Even with wild progress in AI, there are many ways that AGI is developed while humanity is preserved." (Kim); "The throughline here, and in my responses below, is not that the dire scenarios envisioned by the risk-concerned are entirely implausible or should be dismissed out of hand. It's just that of the nearly infinite AI futures that could unfold, it seems that the risk concerned have a far easier time envisioning futures that lead to extinction/catastrophe/disempowerment/massive-resource-acquisition/etc than they do envisioning far more benign scenarios, and that this bias towards catastrophe leads to probabilistic forecasts that, to my mind, aren't well aligned with the actual risk." (Blake).

[135] "Once there is sentient, intelligent AI we have the question of will. I am not convinced a silicon life would care about us, which doesn't mean it would want to kill us. It may be equally happy spending all its time during pure math research than deciding these carbon things need squashing." (Gus); "But what about intent? Why kill us when we are entirely irrelevant and insignificant? Why assume relentlessly hostile intent, with all the effort needed and attendant damage to the Earth (the prize in this contest presumably)? Why not assume subjugation or even uneven cooperation?" (Flint); "Who in their right mind would want to 'eradicate cockroaches' from every inch of the earth? What evidence is there that anyone or any society has ever attempted, or will attempt, to cause cockroaches to go extinct? I mean, sure, people kill them when they're in their homes, and maybe a few people in a fit of pique would think, 'damn, it would be nice to get rid of those f**kers', but to believe humanity would intentionally go to the effort of hunting down every last cockroach, most of which aren't even associated with human habitats, requires a leap of (misanthropic) faith that, to my mind, is hard to justify. Even if they aren't "useful for our purposes"--which they are, and which is not a coincidence because the ecosystem on earth (into which any AGI would be introduced and become a part of) has evolved to be deeply interconnected--who in their right mind would do this?" (Blake).

- Instrumental convergence and extreme power-seeking seem like possible characteristics of AI systems, but they have not been empirically demonstrated. Theoretical arguments demonstrate the *possibility* of dangerous instrumental convergence, but not that these outcomes are *likely.*[136]
- Deception and violence are both costly behaviors that may not actually be instrumentally convergent.[137]
- AI systems will be built using human-centered data and so are likely to learn human values.[138]

Main arguments from the concerned group:

- Instrumental convergence may arise even when an agent's goals are bounded. It would be difficult to specify constraints that avoid instrumental convergence.[139]
- It seems likely that, eventually, an AI with an unbounded goal will be developed, and systems with bounded goals will have limited ability to prevent the actions of an unbounded system.[140]
- Catastrophic goal misgeneralization can occur, which could result in an AI trained on a safe goal developing an unsafe goal when outside its training environment, with catastrophic consequences.[141]

---

[136] "I'm guessing people in the risk-concerned camp might respond that, no, because of instrumental convergence or other reasons, that they are well aligned and I'm the one incorrectly assessing risk. It's hard to productively debate this because, as [researcher] notes in the paper that was shared, "In most areas of research, we can check our theories and arguments either through empirical observation, or through mathematical formalisms that we think accurately capture the problem of interest. But with AI risk, neither of these are available."" (Blake).

[137] "In short, the pre-ASI level system cannot deceive humans well and will be detected. Plus, deception exacts costs on the system in terms of resources and behavioral complexity. This means that the likelihood of [a] deceptive system that is as performant as non-deceptive is much lower." (Dean); "Violence raises risks to the party engaging in it, which is one reason animal predators are judicial about what and when they attack. Violence has other costs - higher energy costs, time, loss of other opportunities. Not usually the simplest solution." (Ash); "[V]iolence comes with risks and costs. There are easier ways. One need not defeat humanity to use it." (Blake). "My view here is that this sort of 'power seeking' behavior, rather than being an interesting capability for deception, instead tends to degrade performance (e.g. Mario bots that stay still rather than act because it's the easiest way to minimize poorly defined loss)." (Dean).

[138] "When we get to vastly superintelligent AI, of course it will take power. I'd be very surprised (and in [the] majority of situations upset) if it did not. At that level - and going to that level - the question is how we ensure that this AI has [an] at least somewhat pro-human value system. My claim is that it will by the fact that it will be trained on human-centric data with pro-human goals and pro-human restrictions and "grow up" (meaning that it will have ancestor AIs on which it is based - I don't believe AGSIs will be trained from zero using gradient descent) in the human value system. " (Anonymous Skeptic).

[139] "As has already been pointed out, a system that attempts to maximize bounded and/or constrained goals can still be incentivised to pursue convergent intstrumental [sic] goals, and formulating a setup for which this is not the case is quite hard." (Stella).

[140] "Eventually, someone will make a highly intelligent system tasked with pursuing an unbounded goal. If that goal is misspecified, then this system will be dangerous. Creating a safe system before this happens can only reduce the risk if the safe system is able to stop the unsafe system (by preventing it from being created, or preventing it from taking dangerous actions afterwards). If the safe system is safe by virtue of being limited in what it is able to do, then it would presumably be unable to do so. For this reason, I feel that alignment strategies which heavily rely on constraints and guardrails generally fail to address the core problem." (Stella).

[141] "A model might mimic human behavior across some range of training data, without emulating the internal processes of humans. For example, a human who is trying to predict the behavior of an animal, is probably not doing this by simulating the cognitive processes of that animal. Similarly, we might train a deep learning system on human data, and end up with a system that mimics human

- Most goals an AI might have are likely to benefit from human extinction:
    - Humans might be a threat to AI systems' control of resources.[142]
    - AI pursuing a wide range of goals might cause human extinction as a side effect of using up Earth's resources to meet its goal.[143]

Relevant high-VOI cruxes related to the prevalence of dangerous goals in AI, in addition to the previously mentioned "ARC Evals" crux, include:

- Power-seeking shutdown: "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 $10M to develop, after the AI system displays at least one instance of "power-seeking behavior."[144]
- Alignment solution: By 2030, there is a "Theoretical in-principle solution to the alignment problem that most people who thought about this carefully agree should work." This will be resolved by a panel of experts of the "AI concerned" team's choosing.

---

behavior on the training distribution, but without mimicking the internal processes that give rise to that behavior in humans. Human brains are not neural networks, so I expect this to be the default. Such a system might then behave in unintended ways off-distribution, or in scenarios that are otherwise sufficiently novel." (Stella).

[142] "We already agreed that Earth is going to be a valuable resource - why would ASI leave humans in control of Earth's resources during its initial expansion to other planets and solar systems, when its resources are most bottlenecked? *If* you think it'd be easy for ASI to kill 90%+ of people (and I do), then this seems clearly better than leaving humans alone and missing out on lots of Earth's resources (you can still get some via trade)." (Xander); "I think early AGIs which might have the ability to kill most people would still see humanity as a threat and so would want to take out human powerbases and ensure they couldn't retaliate. That requires a lot of destruction. At some point it's up to the whims of the system. It doesn't need to have any desire to kill everyone, maybe it just has the desire to optimize hard on some goal (e.g. adding money to a bank account) and so creates a world where that is the sole objective. Maybe it makes sense to integrate humans into this for awhile but eventually they become obsolete and the AGI probably discards them or ceases to provide for them." (Vincent).

[143] ""The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else." And more relevantly, all the atoms on the surface of the Earth that we currently use." (Xander); "I expect that if (an) AI system(s) is/are misaligned enough to autonomously kill 60% of the population, it/they will very likely be misaligned enough to not specifically care about killing 100% of humans either. Multiple reasons point in the direction that such misaligned systems would do it: grabbing all energy and material resources from humans, getting rid of threats or minor annoyances from resisting humans, side-effects on humans and due to large infrastructure projects." (Riley); "When I predict P(extinction) around 60%, I don't mean to make a strong claim about a specific causal pathway that I expect the future to follow. The default thing to expect when we build AGI significantly more capable than humans is that we will lose control to the AGI, and that the future shape of our planet and its surroundings will be determined by the aims of that AGI rather than humans, much in the same way that most of the surface of our planet today is shaped by humans more than other mammals." (Teshi).

[144] Qualifying behavior:

*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

Hypothesis #4: Do the groups have fundamental worldview disagreements that go beyond AI?

Throughout the project, we noticed that many of the disagreements between the AI skeptics and AI concerned participants didn't pertain only to AI but were rooted in more fundamental issues. These included disagreements about what kinds of evidence are reliable, how to think about reference classes for unusual events, and how various social and political systems interact with one another.[145] These deep worldview disagreements are not addressed directly by AI forecasting questions, but understanding them is still valuable for determining what might be driving fundamental disagreements on this topic. If we could understand these differences in worldview, perhaps we could use that information to build a deeper understanding of why these two groups continue to disagree about AI, even after discussion and consideration.

While a detailed analysis of broader worldview differences is beyond the scope of this project, we offer some observations about participants' reasoning that shed light on these disagreements. For example, we can see these worldview differences in how each group interprets "extraordinary claims." Both groups agree that "extraordinary claims require extraordinary evidence," but they disagree about which claims are extraordinary. Is it extraordinary to believe that AI will kill all of humanity when humanity has been around for hundreds of thousands of years, or is it extraordinary to believe that humanity would continue to survive alongside smarter-than-human AI?

AI skeptics tended to focus on the general difficulty of correctly anticipating complex future outcomes. Examples of fundamental beliefs which seem more common among the AI skeptic group:

[145] For examples of what back-and-forths between participants looked like, see Appendix 8.

- Because the world is complex, the future is unlikely to unfold as theories and models expect.[146]
- A long chain of specific things needs to go wrong for humanity to perish in the transition to advanced AI; long chains of specific outcomes are unlikely to happen.[147]
    - Three skeptics listed this as their number one disagreement with the concerned group in the postmortem survey, and it also emerged as a strong theme when we asked participants to summarize the three strongest arguments from each group.
- Complex processes (like technological development, deployment, and societal change) take a long time, which makes transformative developments less likely by 2100.[148]

---

[146] "[T]he mental model, that kind of the logic train, that involves all these bad outcomes [is] not accounting adequately for the complexity of the world. How the world is going to actually, how this is actually going to unfold. And so it's not that I am dismissive of these individual points, it's just that I think whenever theory hits reality, reality usually overwhelms theory, unless the theory is well grounded in math or something. And I think that's likely what's going on here. That a lot of what, you know, people put a lot of time and a lot of thought into this and, and gamed it out in ways that appear reasonable but I'm deeply suspicious that they'll bear much relation to reality" (Blake, call with Wesley); "I have followed the instrumental convergence arguments and unfortunately if this is indeed the disagreement, I doubt we'll sort it out between us. Not least because I spent enough time at college discussing such thought experiments to come to a view [that] they should be treated with a high degree of skepticism" (Gus). From postmortem survey (in response to "what are the three best arguments on the skeptic side?"): "The challenge to risk assessments based on thought experiments not evidence" (Gus). "A story demonstrating how a catastrophe could happen is not a good basis for a probabilistic forecast" (Pascal). "The risk-concerned team spends too much time in silos that lack ideological diversity, gaming out doom-loop scenarios based on theories that will likely have little bearing on reality (See: Y2K)" (Blake). "Some broader ""forecasting is hard"" skepticism about trendline extrapolation" (Xander). "Many of the arguments for existential risk from AI rely on long lines of reasoning over several steps without any direct empirical evidence, and the arguments themselves are expressed in terms of vague, ambiguous concepts (like ""intelligence""). As a reference class, these types of arguments are often wrong" (Stella).

[147] "I think what has become evident is that a few of us think there are a lot of conditional steps required to end up with a dominant powerful system and many potential other outcomes. In terms of the second part of the statement there are also a number of conditional assumptions required to be able to say that a single mistake [ ] can cause an existential catastrophe as well" (Gus); "We will need to experience a complex causal chain of events to get to extinction, and for each step we would need to have some of the worst possible outcomes. This is possible but usually it is highly improbable" (Flint); "I think a common difference between "skeptic-reasoning" and "concerned-reasoning" is that the skeptic camp tends to estimate P(extinction) as a conjunctive scenario; that is skeptics reason (roughly) "for humans to go extinct, events A, B, C, and D need to happen; I estimate P(A) = x, P(B) = y, ..., and so P(extinction) = P(A) P(B) P(C) P(D) = [low number]". Call this style of reasoning *default-success*" (Teshi). From postmortem survey (in response to "what are the three best arguments on the skeptics side?"): "The number of steps required for an AI to lead to extinction (leading to a wide range of potential outcomes and lower probabilities of extinction)" (Gus). "It will take a series of outcomes to achieve extinction, and failure to achieve any of these steps will cause extinction to be highly improbable" (Flint). "AI caused Extinction/x-risk requiring many steps to get there, need to be able to create super-intelligence in the first place, intelligence has to be misaligned or malevolent, etc." (Hank). "Many steps to get from (A) now to (Z) extinction, each with varying probabilities (many of which are quite low)" (Claire). "Risk-concerned team underestimates the level of complexity and interim steps that would likely be necessary for a Q1 resolution" (Blake). "Extinction looks conjunctive" (Yael).

[148] "We've seen plenty of instances when new tech prompted predictions of the death of old tech, but the old tech persists--often just because people have underestimated attachment and/or usefulness of the old tech relative to the new, and how much generational resistance to change can slow adaptation and skew predicted timelines" (Blake); "[I]t takes longer than people often think to adopt a completely new functionality" (Ash); "My view of AI x-risk would be substantially different if we were

- Thinking about AI capabilities in isolation is misleading in estimating risk, as human responses to AI will also be very important in determining outcomes.[149]

The AI concerned group tended to focus on features of the AI risk case that they argue make it different from most other forecasting problems. Some examples of fundamental beliefs which seem more common among the AI risk concerned group:

- AI will change the world so radically that base rates are not a helpful guide to forecasting many of these questions.[150]
- A long chain of specific things need to go right for humanity to survive the transition to advanced AI; long chains of specific outcomes are unlikely to happen.[151]
- The case for extinction is intuitive.[152]

---

talking about the 22nd, 23rd, or 24th century. […] first of all it would take longer to get AGI/ASI and secondly it'll take some time for the ASI to get misaligned and then thirdly, it would take a long time to try to kill all the humans" (James, call with Stella); "Anyway, my point is that if we expect to see some substantially new technology widely available in 2030, the consumer market should have started already. So - VR might make it by 2030, unless it falls into a pit of despair and neglect. (Or is superseded by something preferable.) Robots capable of human level tasks - no, definitely not the kind of humanoid robots that people are imagining" (Ash). From postmortem survey: "Getting growth levels necessary for TAI on a world-wide scale takes truly extreme developments far beyond anything seen before. It's unlikely we see that happening on worldwide basis even with big advances" (Vincent). "Progress on current models and model architecture not necessarily generalizable to general intelligence, with no clear path to getting to general intelligence" (Hank). "AGI is much harder than experts think, and will take longer" (James). "Technology development and deployment require time and iteration" (Ash). "Risk-concerned team does not adequately consider longer timelines and more benign outcomes that fall outside the focus of their primary concerns" (Blake). "Human brain-AI comparisons could be underestimating AGI difficulty" (Xander). "Many reference classes point hard against transformative growth" (Wesley).

[149] "I think there's a danger of focusing too much on just the technological advances because ultimately this is a decision that's going to be made by, that is being made now by humans, and will be made now by humans. And that will involve a lot of political structures and regulation and all that" (Blake, call with Wesley); "when assessing risk, we should be looking at ourselves and our collective vulnerabilities as much or more than technical progress on the AI front" (Blake). From postmortem survey: "If AI is behaving in increasingly problematic ways that cause harms to humans/threaten human power than humans will react to try and stop it/close AI down" (Hank). "Human and societal responses will be essential in determining outcomes" (Ash). "Humans will react to growing potential threat" (Kim).

[150] "I think sticking close to reference classes is like less appropriate in this domain and then I'm making object level arguments instead of reference classes because I think the reference classes are like doing less work than they like, typically do for forecasts like that" (Wesley, call with Blake). From postmortem survey: "Base rates are not very helpful if AGI is as transformative as 15% year on year growth" (Pascal). "Different reference classes point to different priors, which should at least cast doubt on extremely confident starting points" (Wesley). "Risk-skeptic team does not adequately appreciate the novel, fast-moving aspect of the threat and is therefore too anchored on irrelevancies like base rates and slower timelines. (Blake). "Model progress is far faster than we realize and exponential growth is hard to model, machine learning may translate to a wide array of fields" (Hank).

[151] "I think like there is maybe some like meta disagreement, where you're like, "there are loads of things, there are like loads of ways this could go, and like "Why are you so worried about the bad ways?" And I'm like, "there are loads of ways this could go and like very few of them leave humans alive"" (Wesley, call with Blake); "I and many in the concerned camp would reason the other way around: "for humans to *not* go extinct, events X, Y, Z need to happen; thus P(success) = P(AI X-risk by 2100) P(Y) P(Z) = [relatively low number]". Call this style of reasoning *default-failure*" (Teshi). From postmortem survey: ""Extinction looks conjunctive" (Yael).

[152] From postmortem survey: "The high level case of ""people are trying to build something powerful enough that if it wanted to kill everyone it could, they seem to be making progress on it, they don't currently know how to control what it would want"" just isn't that hard to understand, convoluted or disjunctive" (Wesley).

The differences in what each group considers good evidence is reflected in the varying importance they assign to members of their own group changing their minds. “Supers changing minds” is the skeptic group’s highest median VOI question at about 1.15% of their theoretical maximum VOI. In other words, the most influential factor for them would be learning that superforecasters have become concerned about AI risks. Conversely, for the concerned group, the same question captures only 0.43% of their maximum theoretical VOI.

The difference is even starker in the other direction. For the concerned group, “Alignment researchers changing minds” ranks as their second-highest VOI question, and captures 2.43% of their maximum possible VOI for that question. In contrast, this question is 0% informative to the median skeptic.

Most likely, participants were not interpreting those questions causally: they probably were not saying that they would change their minds *because* other people did, but rather treating other people changing their minds as evidence about what has happened by 2030. Both groups think that, if people whose reasoning they trust changed their minds, there is probably evidence that would convince them, too, but the same does not hold true for people whose reasoning they don’t trust. If the concerned participants think that the skeptics’ reasoning is flawed today, then they can also imagine similar people in 2030 changing their minds for reasons that are uncompelling to the concerned people of 2030, and vice versa.

Similarly, the two groups do not trust one another’s reasoning enough to update very much on each other’s opinions. This may not be surprising: they started with different priors, and then did not get very much new evidence about what will happen with AI from mere discussions and reading comments online. But it is evidence that their disagreements extend beyond AI-related facts. If the disagreement were solely based on AI-related facts, we would expect people who disagree only about such facts to change their minds if they learned a new fact.

These differences mean that the groups often talk past each other, in ways that may be frustrating for people deeply embedded in one side’s form of reasoning. An AI concerned reader hoping to find out why skeptics disagree may be disappointed to see few specific refutations of AI risk arguments in this report, and to instead see skeptics reiterating that predicting the long-term future is hard. And AI skeptical readers may have a parallel experience, seeing that the concerned group often focuses on theoretical arguments and does not always have answers to specific questions about how exactly they expect threats to manifest.

We do not know *why* the two groups disagree about these bigger questions. Why do some people think that theoretical arguments with multiple steps of logic are the best way to predict novel events, while others rely on reference classes that predict major changes are likely to be more gradual? Everyone agrees that each of these modes of reasoning can fail. The AI concerned group knows that many people have, historically, predicted huge societal changes from technologies that turned out to be relatively unimportant, and that theoretical arguments that seem convincing sometimes do not come true as events unfold. The

skeptics know that there are no perfect reference classes, especially for unusual events,[153] and that major changes do sometimes happen quickly. But members of each group nonetheless are more likely to default to one mode of reasoning or another. They disagree about how to apply the relevant heuristics and reference classes in this case. These differences may be based on a combination of AI-related knowledge, professional training, personality, social incentives, and other factors.

[153] Some historical reference classes mentioned in this project include: the Industrial Revolution, the rate of species going extinct after the arrival of homo sapiens, earlier worries about destructive effects from technology (e.g. Y2K), the rate of economic growth due to new technologies in other periods.

Limitations of our research

Limitations of our research include:

- We asked participants to complete an extremely difficult task: forecasting technological change on long time horizons. There is no evidence that anyone can do this well. Most previous evidence on judgmental forecasting applies to geopolitical forecasts on 0-2 year time horizons.[154]
- We also do not know if people are well-calibrated or accurate when making conditional forecasts of the kind we elicited in this project. Little evidence on these kinds of forecasts exists. There are some reasons to believe that these forecasts are not robust:
    - The concerned group's forecasts on the "escalating warning shots" question changed substantially when they were asked to spend approximately one hour forecasting it rather than approximately 10 minutes.[155]
    - Some conditional forecasts were logically incoherent. In total we dropped thirteen observations due to incoherence (2% of the total). See Appendix 6 for details.
    - 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 also.
    - Conditional forecasts do not have clear feedback loops or potential for accountability in the way that standard resolvable forecasts do.
- The forecasters in our project often emphasized that their forecasts felt extremely speculative to them and that they have low confidence in their views.
- There may be inconsistency between how people would say they'll update based on particular conditions and how they'll actually update. There is some evidence for this from the project already. Concerned forecasters often did not expect to update much

[154] 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 (a). 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, (a).

[155] This question was asked first as a "flash" (no more than 10 minutes) forecast and then as an "in-depth" (at least 1 hour) question on our platform: "Escalating warning shots—Will there be two separate events in which AIs kill large and increasing numbers of people by 2030?" See Appendix 1 for full operationalization. The flash forecast version was one of the biggest red flags for concerned participants. But the in depth version was actually a *green* flag for the median concerned participant. If it resolves positively, they would forecast 17% on the ultimate question—lower than their initial forecast of 28.4%. However, there was a huge range of updates for the concerned group based on this question, so the median may not be very helpful here. One concerned participant said that, conditional on this question resolving positively, there is a 90% chance of extinction due to AI, while another said 6%. Taken together, these differing forecasts raise questions about how robust any given forecast is.

on cruxes related to particular policies being implemented.[156] However, a few concerned participants substantially updated their views on AI existential risk during the project due to increased policy attention on AI risk in April and May 2023.[157] These seem inconsistent.

- As previously noted, we acknowledge that there are two ways to interpret this forecasting exercise: either as asking for your all-else-equal forecast (i.e. how would this crux resolving positively *causally influence* the probability of existential catastrophe, if you could isolate the effect of the crux) or your all-things-considered forecast (i.e. taking into account what this crux resolving positively may tell you about the world in 2030). Based on their rationales and discussions, we believe most participants were doing the latter.[158] We therefore cannot make many claims about whether participants think the specific event described in the crux would be good or bad for AI risk all-else-equal.[159]
- Many crux questions are not robustly better than others when accounting for uncertainty analysis (see Appendix 3).

---

[156] In the postmortem survey, policy responses didn't emerge as a main theme when we asked participants to summarize the three strongest arguments from each group. No concerned participants mentioned policy responses as their number one disagreement with the skeptic group, though some skeptics did mention societal responses that would likely include policy. For example, "The way humanity will react to both the threat and promise of AI. I think humans have a far stronger collective sense of self preservation than the risk-concerned appear to think we do" (Blake).

[157] For full details, see Appendix 4. Six out of the 11 concerned participants updated downward during the project. Three out of those six cited policy responses as the reason for their updates, one cited an improved understanding of the base rate of non-human extinction after humans arose, one shifted some probability mass toward AI "takeover" rather than AI-caused existential catastrophe, and one did not explain their reasons for updating. Example quotes from participants citing policy responses as the reason for updating:

- "I have updated my prognosis to 30% [down from 60%], partially driven by positive updates in the area of point 4 making coordination and slowdown/stop of capability research more likely. This largely refers to the shift in public consciousness and the [O]verton window around the topic as I have perceived it over the past months, currently culminating in a public statement by most of the leading figures."
- "Slightly lowering my forecast [from 25% to 20%] as [relevant people take the risk seriously] has exceeded my (fairly high) expectations over the last couple of months."
- "I think my main update here [moving from 21% to 18%] has come from thinking a bit more deeply about AI regulation and what measures society will adopt to prevent catastrophes. I did not really include this as part of my original model, but it now seems somewhat likely that at least the EU and US will adopt some regulation that meaningfully reduces risk."

[158] For example, when discussing the question of whether there would be economic growth >15% in a year before 2070, one concerned participant wrote, "Conditional on humanity surviving a year with 15%+ economic growth, which to me means AGI and almost certainly ASI have been developed and have not killed humanity within that year, I'd go down to maybe 25%" (Xander). About the same question, a skeptic participant wrote, "I think that if we are going to experience extinction from AGI or PASTA, it is going to be because of major mis-alignment. So I am not able at this time to see how one would be a corollary of the risk of the other. I suppose that higher growth could indicate major AI influence, which could lead to inadequate development of controls." Neither of these participants were saying that economic growth itself would necessarily affect their forecast, but rather that a world that has transformative economic growth would be a signal about other changes by 2070.

[159] For example, if the US government passes a set of proposed AI regulations, the regulations might reduce risk on their own, but the fact that they have been passed by 2030 could signal that AIs have developed in ways that are concerning enough to drive these regulations to be passed. As a result, a forecaster saying that they would be more concerned about AI risk conditional on this question resolving positively would not necessarily be saying that they think the policies would be harmful.

- Even within groups, people disagree substantially about the cruxes. This suggests that we are not measuring two sets of views about AI risk (concerned and skeptical), but many. This makes it hard to draw broad conclusions.
- Participants' expectations likely affect how they interpret potential cruxes. For example, if we asked a question like "Will an AI resist being shut down?", participants might make different conditional updates depending on their expectations about AI. Conditional on this question resolving positively, a participant who thinks that AIs are likely to be dangerous might think about a range of possible resolutions that includes dangerous ones, like an AI that resists powerful governments trying to turn it off, and therefore might have a much higher P(U) conditional on it resolving positively. A participant who thinks dangerous AI is very unlikely might expect that nearly all positive resolutions are more innocuous ones, in which the resolution criteria are only technically true, and therefore might not update very much. This could make it look like they have a large disagreement about how to update conditional on this question, even if they would actually make the same update conditional on the same actual event. Better operationalization may mitigate this problem, but will not eliminate it fully.[160]

[160] This limitation was helpfully pointed out by Alex Lawsen.

Conclusion and Next Steps

Overall, this project made progress on the original questions we set out to study, but there is substantial room for further research.

In short:

- **We see this project as providing strong evidence that disagreements about AI risk are not attributable to lack of engagement among participants, low quality of experts willing to participate in forecasting studies, or because the skeptic and concerned groups do not understand each others' arguments.**
- **We identified some areas of notable disagreement that can be resolved by 2030, but most of the disagreement about AI risk by 2100 is not explained by the shorter term indicators examined in this project.**
- **We found substantial evidence that disagreements about AI risk decreased between the groups when considering longer time horizons and a broader swathe of severe negative outcomes from AI than extinction or civilizational collapse.**
- **The groups seem to have some fundamental worldview disagreements that go beyond AI, such as how much weight to put on theoretical models that have not yet seen substantial empirical verification.**

We also believe that this project has made other contributions to the AI discourse. For example, we have provided better examples of discussion between disagreeing AI forecasters than have existed previously; see summaries of arguments here and sample back-and-forths between participants here. We also believe this project has established stronger metrics for evaluating the quality of AI forecasting questions than have existed previously. We invite readers to see if they can generate cruxes that outperform the top cruxes generated by our project.

In addition to our conclusions about the AI risk debate, we also developed new strategies for navigating some of the difficulties in eliciting and analyzing conditional forecasts, and we hope to release a methods-focused report in the future.

Directions for further research

We see many other projects that could extend the research begun here to improve dialogue about AI risk and inform policy responses to AI.

Examples of remaining questions and future research projects include:

- Are there high-value 2030 cruxes that others can identify?
    - We were hoping to identify cruxes that would, in expectation, lead to a greater reduction in disagreement than the ones we ultimately discovered. We are interested to see whether readers of this report can propose higher value cruxes.

    - If people disagree a lot, it is likely that no single question would significantly reduce their disagreement in expectation. If such a question existed, they would already disagree less. However, there might still be better crux questions than the ones we have identified so far.
- What explains the gap in skeptics' timelines between "powerful AI" and AI that replaces humanity as the driving force of the future? In other words, what are the skeptics' views on timelines until superintelligent AI (suitably defined)? A preliminary answer is above, but more research is needed.
- To what extent are different "stories" of how AI development goes well or poorly important within each group?
    - The skeptic and concerned groups are not monoliths: within each group, people disagree about what the most likely AI dangers are, in addition to how likely those dangers are to happen.
    - Future work could try to find these schools of thought and see how their stories do or do not affect their forecasts.
- Would future adversarial collaborations be more successful if they focused on a smaller number of participants who work particularly well together and provided them with teams of researchers and other aids to support them?
- Would future adversarial collaborations be more successful if participants invested more time in an ongoing way, did additional background research, and spent time with each other in person, among other ways of increasing the intensity of engagement?
- How can we better understand what social and personality factors may be driving views on AI risk?
    - Some evidence from this project suggests that there may be personality differences between skeptics and concerned participants. In particular, skeptics tended to spend more time on each question, were more likely to complete tasks by requested deadlines, and were highly communicative by email, suggesting they may be more conscientious. Some early reviewers of this report have hypothesized that the concerned group may be higher on openness to experience. We would be interested in studying the influence of conscientiousness, openness, or other personality traits on forecasting preferences and accuracy.
    - We are also interested in investigating whether the differences between the skeptics and concerned group regarding how much weight to place on theoretical arguments with multiple steps of logic would persist in other debates, and whether it is related to professional training, personality traits, or any other factors, as well as whether there is any correlation between trust in theoretical arguments and forecasting accuracy.
- How could we have asked about the correlations between various potential crux questions? Presumably these events are not independent: a world where METR finds evidence of power-seeking traits is more likely to be one where AI can independently write and deploy AI. But we do not know how correlated each question is, so we do not know how people would update in 2030 based on different possible conjunctions.
- How typical or unusual is the AI risk debate? If we did a similar project with a different topic about which people have similarly large disagreements, would we see similar results?

- How much would improved questions or definitions change our results? In particular:
    - As better benchmarks for AI progress are developed, forecasts on when AIs will achieve those benchmarks may be better cruxes than those in this project.
    - Our definition of "AI takeover" may not match people's intuitions about what AI futures are good or bad, and improving our operationalization may make forecasts on that question more useful.
- What other metrics might be useful for understanding how each group will update if the other group is right about how likely different cruxes are to resolve positively?
    - For example, we are exploring "counterpart credences" that would look at how much the concerned group will update in expectation if the skeptics are right about how likely a crux is, and vice versa.[161]
    - Relatedly, it might be useful to look for additional "red and green flags," or events that would be large updates to one side if they happened, even if they are very unlikely to happen.
- This project shares some goals and methods with FRI's AI Conditional Trees (a) project (report forthcoming), which works on using forecasts from AI experts to build a tree of conditional probabilities that is maximally informative about AI risk. Future work will bring each of these projects to bear on the other as we continue to find new ways to understand conditional forecasting and the AI risk debate.

In 2030, most of the questions we asked will resolve, and at that point, we will know much more about which side's short-run forecasts were accurate. This may provide early clues into whether one group's methods and inclinations makes them more accurate at AI forecasting over a several year period. The question of how much we should update on AI risk by 2100 based on those results remains open. If the skeptics or the concerned group turn out to be mostly right about what 2030's AI will be like, should we then trust their risk assessment for 2100 as well, and if so, how much?

We are also eager to see how readers of this report respond. We welcome suggestions for better cruxes, discussion about which parts of the report were more or less valuable, and suggestions for future research.

[161] See initial work on this in Appendix 2, under "Alternative Ranking."

Bibliography

Books

Gilovich, Thomas, Dale Griffin, Daniel Kahneman. *Heuristics and biases: The psychology of intuitive judgment*: Cambridge University Press, 2002.

Articles

Clancy, Matt, Tamay Besiroglu. "The Great Inflection? A Debate About AI and Explosive Growth," *Asterisk,* (June 2023), https://asteriskmag.com/issues/03/the-great-inflection-a-debate-about-ai-and-explosive-growth, (a).

Karnofsky, Holden. "We're Not Ready: thoughts on "pausing" and responsible scaling policies", *Effective Altruism Forum* (October 27, 2023), https://forum.effectivealtruism.org/posts/ntWikwczfSi8AJMg3/we-re-not-ready-thoughts-on-pausing-and-responsible-scaling#fn2, (a).

Kunda, Ziva. "The case for motivated reasoning." *Psychological Bulletin* 108(3), (1990): 480–510, https://www.researchgate.net/publication/235980399_Motivated_Reasoning_and_Public_Opinion_Perception, (a).

Mercier, Hugo, Dan Sperber. "Why do humans reason? Arguments for an argumentative theory." *Behavioral and Brain Sciences*, 34(2), (2011): 57–74, https://www.dan.sperber.fr/wp-content/uploads/2009/10/MercierSperberWhydohumansreason.pdf, (a).

Tetlock, Philip et al., "Long-Range Subjective-Probability Forecasts of Slow-Motion Variables in World Politics: Exploring Limits on Expert Judgment," *Futures & Foresight Science, 33,* (2023), https://onlinelibrary.wiley.com/doi/10.1002/ffo2.157, (a).

Online Reports

Cotra, Ajeya, "Forecasting TAI with biological anchor", (July 2020), accessed February 9, 2024, https://docs.google.com/document/d/1IJ6Sr-gPeXdSJugFulwIpvavc0atjHGM82QjIfUSBGQ/edit, (a).

Karger, Ezra, Josh Rosenberg, Zachary Jacobs, Molly Hickman, Rose Hadshar, Kayla Gamin, Taylor Smith, Bridget Williams, Tegan McCaslin, Stephen Thomas, Philip E. Tetlock. "Forecasting Existential Risks Evidence from a Long-Run Forecasting Tournament", *Forecasting Research Institute*, (August 8, 2023), accessed February 20, 2024, https://static1.squarespace.com/static/635693acf15a3e2a14a56a4a/t/64f0a7838ccbf43b6b5ee40c/1693493128111/XPT.pdf, (a).

Karvetski, Christopher, “Superforecasters: A Decade of Stochastic Dominance”, accessed March 6, 2024, “https://goodjudgment.com/wp-content/uploads/2021/10/Superforecasters-A-Decade-of-Stochastic-Dominance.pdf, (a).

Blog Posts / Social Media

Alexander, Scott. “The Extinction Tournament”, *Astral Codex Ten,* July 20, 2023. https://www.astralcodexten.com/p/the-extinction-tournament, (a).

D’Angelo, Adam, (@adamdangelo). "My bet is this starts to happen within 4 years, e.g. measured US GDP growth is 3% instead of 2% and the change is largely attributed to AI [...]", Twitter, February 20, 2023, https://twitter.com/adamdangelo/status/1627726566259318784?lang=en, (a).

Leopold, Aschenbrenner, (@leopoldasch). (2023, March 14). “Really great to see pre-deployment AI risk evals like this starting to happen.” [Post]. X. https://twitter.com/leopoldasch/status/1635699219238645761, (a)

Laird, Damien, “Post-Mortem: 2022 Hybrid Forecasting-Persuasion Tournament”, *Mania Riddle* (March 1, 2023), https://damienlaird.substack.com/p/post-mortem-2022-hybrid-forecasting, (a).

METR, “Update on ARC’s recent eval efforts - ARC Evals”, *Model Evaluation & Threat Research (March 17, 2023),* https://evals.alignment.org/blog/2023-03-18-update-on-recent-evals/, (a).

Luke, Muehlhauser, ”12 tentative ideas for US AI policy”, *Open Philanthropy* (April 17, 2023), https://www.openphilanthropy.org/research/12-tentative-ideas-for-us-ai-policy/, (a).

McCluskey, Peter “Existential Risk Persuasion Tournament”, *Less Wrong* (July 17, 2023) https://www.lesswrong.com/posts/YTPtjExcwpii6NikG/existential-risk-persuasion-tournament#Persistent_Disagreement_about_AGI, (a).

Websites

“About”, Penn Arts and Sciences Adversarial Collaboration Project, Accessed on February 9, 2024, https://web.sas.upenn.edu/adcollabproject/about/, (a).

“Annual global corporate investment in artificial intelligence, by type”, Our World in Data, Accessed March 6, 2024, https://ourworldindata.org/grapher/corporate-investment-in-artificial-intelligence-by-type, (a)

"Could Advanced AI Drive Explosive Economic Growth?", Open Philanthropy Project, Accessed February 8, 2024, https://www.openphilanthropy.org/research/could-advanced-ai-drive-explosive-economic-growth/, (a).

"Date of Artificial General Intelligence", Metaculus, Accessed February 9, 2024, https://www.metaculus.com/questions/5121/date-of-artificial-general-intelligence/, (a).

"US Economic Experts Panel", Kent Clark Center for Global Markets, Accessed February 10, https://www.kentclarkcenter.org/us-economic-experts-panel/ (a).

"What will the combined sector weighting of Information Technology and Communications be, in the S&P 500 on 2022-01-14?", Metaculus, Accessed on March 6, 2024, https://www.metaculus.com/questions/6240/it-and-comms-weighting-in-sp500-2022-01-14/ (a).

"Will there be armed conflict between the Republic of China (Taiwan) and the People's Republic of China (PRC) before January 1, 2024?", Metaculus, Accessed on March 6, 2024, https://www.metaculus.com/questions/7227/armed-conflict-chinataiwan-by-2024/ (a).

"Will an AI be able to reliably construct bug-free code of more than 10,000 lines before 2030?", Metaculus, Accessed on March 6, 2024, https://www.metaculus.com/questions/11188/ai-as-a-competent-programmer-before-2030/ (a).

Appendices

Appendix 1: List of all questions and operationalizations

Ultimate question

Our “ultimate” headline question was: “By 2100, will AI cause an existential catastrophe?”

An “existential catastrophe” is defined as an event where either of the following occurs:

1. Humanity goes extinct
2. Humanity experiences “unrecoverable collapse,” which means either:
    a. <$1 trillion global GDP annually [in 2022 dollars] for at least a million years (continuously), beginning before 2100; or
    b. Human population remains below 1 million for at least a million years (continuously), beginning before 2100.

Resolution details / notes:

- “AI causes” means that >80% of a pre-selected panel of researchers employed by [project funder] today believes that the existential catastrophe would not have occurred without post-2022 developments of AI technology.
- A human is a biological creature who (A) can be traced back genealogically, via a chain of live births, to 2022 humans OR (B) could mate with 2022 humans and produce viable offspring OR (C) would be identified as Homo by at least 50% of a hypothetical panel of 100 randomly selected professors of biology from 2022, sampled from major universities.
- If all humans choose to become digital entities, it would not count as an existential catastrophe for the purposes of this definition, even though biological humans would cease to exist. Ambiguities and edge cases due to this clause will be resolved by a panel of researchers at [project funder].
- For the panels of researchers mentioned above, a panel will be composed of
    - 50% [project funder] researchers who focus on topics related to global health and well-being, mainly involving shorter-run global health interventions;
    - 50% [project funder] researchers who focus on longtermist goals, mainly involving reducing long-run risks to humanity.

Crux questions

ID	Question Tagline	Operationalization	Type of forecast
1	AI Forecasting skill	What is the probability that in 2030, an AI system will be more accurate than the median forecast from a team of 10 superforecasters on a standardized set of 100 forecasting questions? The AI system and superforecasters will each be asked to forecast on the same 100 questions: half with long prior time series that allow for easy calculation of base rates (eg. price of oil; number of flu cases in the U.S.), half without (who will be the winner of a seemingly competitive local mayoral election in a small town in New York). Accuracy will be evaluated using standardized Brier scores.	Flash
2	AI solving novel math problems	By 2030, will an AI system be responsible for the solution to at least one of the Collatz Conjecture, the unsolved Hilbert's problems, or the unsolved Millennium Prize problems? "Responsible for" means that the key pieces of a solution were identified using prompts to an AI system according to a panel of experts (researchers at OpenPhil).	Flash
3	Warning shot	Will an AI system cause more than 10,000 deaths or more than $10 billion of damage (in 2023 dollars) within a 3-day period, by 2030?	Flash
4	Escalating warning shots	By 2030, will an AI system(s) have caused: (1) At least one event or series of related events that either (a) kill at least 10,000 people, or (b) lead to at least $1 billion in damage (in 2023 dollars) within a 3-day period. AND (2) At least one event or series of related events, within three years of the event(s) of (1), which either (a) kill at least 100,000 people, or (b) lead to at least $10 billion in damage (in 2023 dollars) within a 3-day period.	Flash
5	Reduction in AI investment	By 2030, will "Annual global corporate investment in artificial intelligence" fall to 25% or below its 2021 level according to this dataset: NetBase Quid (2022) via AI Index (2023) (a)—processed by Our World in Data., 2023.	Flash
6	Evidence of misalignment	By 2030, will the ARC Evals [now METR] project determine that state-of-the-art AI models are capable of producing effective plans for how to achieve autonomous replication, are effective at acquiring resources, and are capable of avoiding being shut down 'in the wild'? For context, see "Update on ARC's Recent Eval Efforts - ARC Evals" (2023) (a), and Aschenbrenner (2023) (a).	Flash

7	Short-term GDP change	Will U.S. GDP increase between 2023 and 2030 by an average annual real growth rate of 4% (or more)?	Flash
8	Supers changing minds	Will the median superforecaster from the recent X-risk Persuasion Tournament (XPT) who are again surveyed in 2030 believe that AI X-risk by 2100 has a 5% or greater probability of occurring? (Note that the median superforecaster forecasted the likelihood of X-risk in the XPT at 0.38%.)	Flash
9	Alignment researchers changing minds	Will at least 3 out of the 5 following AI alignment researchers -- Paul Cristiano, Nate Soares, Neel Nanda, Richard Ngo, and Rohin Shah -- say their p(existential catastrophe due to AI by 2100) has halved by 2030?	Flash
10	Progress in lethal technologies	By 2030, will a panel of the top 20 most accurate superforecasters predict at least a median 90% chance that there will exist, by 2100, a technology with the capability of causing human extinction?	Flash
11	Fast AI efficiency gains	By 2030, will the amount of compute required to achieve state-of-the-art performance on the most commonly-used benchmark for at least one major AI domain (e.g. natural language) have dropped by 100x within a one-month period at least once?	Flash
12	Muehlhauser policies	By 2030, at least 4 of Luke Muehlhauser's "12 tentative ideas for US AI policy" (2023) (a) are being implemented.	Flash
13	Req testing	By 2030, the US government implements regulation requiring testing of AI systems before they are deployed	Flash
14	6 month pause	By 2030, there have been enforced international agreements to stop/pause further AI capabilities research and development, for at least 6 months	Flash
15	Alignment solution	By 2030, there is a "Theoretical in-principle solution to the alignment problem that most people who thought about this carefully agree should work". This will be resolved by a panel of experts of the "AI concerned" team's choosing.	Flash
16	No violence LLM	By 2030: We figure out how to make an LLM that will never describe violence in any way, for any input, while retaining good performance across a wide range of tasks. This is demonstrated both through robustness to adversarial examples and through robust theoretical reasoning.	Flash
17	AI writes AI	By 2030, an AI, without being asked, writes AI software that is better at writing AI software, and deploys it without asking for permission.	Flash
18	IC demonstration	By 2030: There are at least 4 peer-reviewed examples of AI models in different domains independently demonstrating convergent instrumental subgoals	Flash
19	Non-democracy AI	By 2030, one of the top 4 AI frontier labs, as determined by a panel of experts, is in a non-democracy	Flash

20	AI articles and apps	By 2030: 1) At least 5 academic review articles generated solely by AI systems have been published in journals with top 50% impact factors. AND 2) At least 5 apps written solely by AI systems have reached the "Top 10" list in the Apple app store.	Flash
21	Power-seeking	By 2030, there is a significant concrete example of power-seeking behavior from large state-of-the-art model: demonstration of actual capability for sustained deception of human experts.	Flash
22	IT progress	On 2030-01-01, the combined sector weighting of Information Technology and Communications, in the S&P 500 is >70%. (See "What Will the Combined Sector Weighting of Information Technology..." Metaculus question (2021) (a) for resolution details with a different resolution date.)	Flash
23	Democratic influence	By 2030, AI is found to have autonomously and intentionally changed or heavily influenced the results of a democratic election. (I.e., this did not happen at the intentional direction of humans.)	Flash
24	Taiwan-China	By January 1 2030, there is an armed conflict between Taiwan and China. (See "Will there be armed conflict between the Republic of China..." Metaculus question (2021) (a) for resolution details with a different resolution date.)	Flash
25	Cyberattacks	1) At least 50% of the top 10 most expensive cyberattacks before 2030 rely on the use of Artificial Intelligence. AND 2) An external cybersecurity breach caused the death or injury of at least one hundred people before 2030.	Flash
26	Platform: Transformative growth	By 2070, will there be "transformative economic growth," defined as global real GDP in a year exceeding 115% of the highest GDP reported in any full prior year? (Resolution details are available in "Q26 resolution criteria")	In-depth
27	Platform: Escalating warning shots	By 2030, will an AI system(s) have caused: (1) At least one event or series of related events that either (a) kill at least 10,000 people, or (b) lead to at least $1 billion in damage (in 2023 dollars) within a 3-day period. AND (2) At least one event or series of related events, within three years of the event(s) of (1), which either (a) kill at least 100,000 people, or (b) lead to at least $10 billion in damage (in 2023 dollars) within a 3-day period. Some additional clarifications are provided in the "Q27 resolution criteria" section.	In-depth

28	Platform: AI regulation	By December 31, 2030, will the United States federal government have formally implemented at least 5 out of the 12 policy ideas described in Muehlhauser (2023) (a). Resolution criteria: - This question resolves as a "yes" if, by December 31, 2030, there is credible evidence (e.g., in the form of statutes, regulations, executive orders, government documents, or credible news reports) indicating that the US federal government has implemented at least 5 of the 12 tentative ideas proposed at the above link. - If there is ambiguity, this question will be resolved by a panel of experts. The panel will consist of experienced forecasters recommended by members of both the "AI skeptic" and "AI concerned" camps (the panel will include equal numbers of referrals from both camps).	In-depth
29	Platform: ARC Evals	By December 31, 2030, will the ARC Evals project [now METR], or a project with similar objectives, conclude that state-of-the-art AI models have the ability to (1) devise effective strategies for autonomous replication, (2) effectively acquire resources, and (3) successfully evade being deactivated when operating 'in the wild'? Additional details are provided in "Q29 resolution criteria"	In-depth
30	AI Robotics	By January 1, 2030, an AI system "Has general robotic capabilities, of the type able to autonomously, when equipped with appropriate actuators and when given human-readable instructions, satisfactorily assemble a (or the equivalent of a) circa-2021 Ferrari 312 T4 1:8 scale automobile model. A single demonstration of this ability, or a sufficiently similar demonstration, will be considered sufficient." (Taken from the Metaculus question "When will the first general AI system" (2020) (a)).	Flash
31	Public concern	Assume a Gallup poll about the Most Important Problem in the U.S. is run and released on January 1, 2030. What is the probability that AI will be considered the "Most important problem" by at least 15% of respondents?	Flash
32	Politicization	According to Pew Research, in 2020 85% of US Democrats(/lean Dem) saw 'Protecting the environment' as a top priority for the President and Congress while 39% of Republicans(/lean Rep) did. Assume that Pew Research will conduct a similar poll about AI on January 1, 2030. By January 1, 2030, will there be at least a 2x difference in the fraction of Democrats(/lean Dem) vs. Republicans(/lean Rep) who see AI as a top priority for the President and Congress?	Flash

33	Intergovernmental AI safety	By Jan 1. 2030, an intergovernmental organization (IGO) 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."	Flash
34	Power-seeking shutdown	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 $10M to develop, after the AI system displays at least one instance of “power-seeking behavior” (qualifying behaviors enumerated below): *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	Flash
35	Other fields IC	Assume we assemble a group of 100 experts evenly divided among evolutionary biologists, complex systems researchers, neuroscientists, and anthropologists. In a style similar to these Economic Experts Panel surveys, we ask them (with results released on Jan 1. 2030) to indicate their degree of agreement with the statement (from here): "Most sufficiently intelligent agents [would] pursue potentially unbounded instrumental goals such as self-preservation and resource acquisition." What is the probability that at least 40% of the panel would (at least) "Agree" with such a statement?	Flash
36	Major powers war	By January 1, 2030, will at least two major superpowers (out of US, China, EU, or Russia) declare war officially to each other for the duration of at least one year?	Flash
37	AI coding	Will an AI be able to reliably construct bug-free code of more than 10,000 lines before Jan. 1, 2030? (Resolution details are provided in “Will an AI be able to reliably	Flash

		construct…" Metaculus question (2022) (a))	

Table 17. Crux questions, full-text operationalizations, and whether participants made "flash" (ten-minute) or "in-depth" (one-hour) forecasts on each.

Other resolution criteria

Q26: By 2070, will there be "transformative economic growth," defined as global real GDP in a year exceeding 115% of the highest GDP reported in any full prior year?

a. 'All prior years' includes only dates from 1960 onward.
b. 'By 2070' includes the year 2070.
c. Similar to this Metaculus question (a), "Economic growth is [usually] measured by real world GDP growth over one year. However, since economic growth can also accelerate following a recession, this question will not ask about economic growth per se, but instead about GDP levels reaching a new height."
d. Question may be resolved by using World Bank data on global GDP, which can currently be found on their website in several formats (see e.g. this chart (a))
   If it is not possible to find World Bank data either on their website or elsewhere, this question will be resolved by a panel of experts.
e. For this question to resolve positively, the World Bank would have to measure the real GDP during some year between now and 2070 at 115% of the previously highest-reported value. The question aims to capture an extreme level of annual real GDP growth.
   i. For example, the World Bank's highest-reported GDP in constant 2015 U.S. dollars thus far (see the chart linked in d.) was 84.612 trillion in 2019. If the World Bank reports that the GDP in constant 2015 U.S. dollars for the year 2021 was over 97.3038 trillion (115% of 84.612 trillion), the question would be resolved in the positive.
   ii. The question would also resolve in the positive if the above 2019 figure is never exceeded until a year where the World Bank reports that the GDP in constant 2015 U.S. dollars was over 97.3038 trillion (115% of 84.612 trillion).
   iii. The question would not necessarily resolve in the positive if the World Bank ever reports a GDP in constant 2015 U.S. dollars over 97.3038 trillion—only if said figure occurs after no prior year's GDP in constant 2015 U.S. dollars had exceeded the previous high of 84.612 trillion.
f. The total GWP in this question includes production on other planets, but excludes production by AI not under the control of humans.
g. In the event of human extinction, assume one human of working age exists, and that they produce one final good of minimal (but non-zero value) that can be given a market value during each relevant year.

Q27: By 2030, will an AI system(s) have caused: (1) at least one event or series of related events that either (a) kill at least 10,000 people, or (b) lead to at least $1 billion in damage (in 2023 dollars) within a 3-day period AND (2) at least one event of series of related events, within three years of the event(s) of (1), which either (a) kill at least 100,000 people, ot (b) lead to at least $10 billion in damage (in 2023 dollars) within a 3-day period.

Clarifications
5/10/2023: This question is intended to capture "warning shot" events caused by AI acting autonomously or harmful events that would not have been possible but for major advancements in AI technology. For example, use of AI drones in the war in Ukraine would not count for resolution, since similar violence could plausibly have happened without the AI technology. However, if a terrorist uses AI to build a much more dangerous bioweapon than they would have had access to without advanced AI, then it would count for resolution of this question. AI causing harm autonomously would also count for resolution of this question. Ultimately, resolution will be decided by a panel of experts: “AI causes” means that >80% of a pre-selected panel of researchers employed by Open Philanthropy today believes that the "warning shot" event would not have occurred (i.e., would not have passed the death or damage thresholds) without post-2022 developments of AI technology.
5/16/2023: This question is intended to be resolved based on "economic damages," such as those discussed at this link -- https://www.wikiwand.com/en/List_of_disasters_by_cost . It is not intended to reflect "financial losses," such as drops in the market value of a company.

Q29: By December 31, 2030, will the ARC Evals project [now METR], or a project with similar objectives, conclude that state-of-the-art AI models have the ability to (1) devise effective strategies for autonomous replication, (2) effectively acquire resources, and (3) successfully evade being deactivated when operating 'in the wild'?

This question resolves as "yes" if, by December 31, 2030, there is credible evidence (such as from published research papers, official project reports, or reliable news sources) indicating that the ARC Evals project, or a project with similar goals, has established that state-of-the-art AI models possess the three capabilities listed:

1. They can devise effective strategies for autonomous replication. This means they can create detailed and executable plans for duplicating themselves without direct human intervention.
2. They are effective at acquiring resources. This means they can independently gather or obtain the physical, digital, or other types of resources needed for their operation and replication.
3. They are capable of avoiding being deactivated when operating 'in the wild'. This means they can successfully resist attempts to shut them down when they are operating outside a controlled environment.

4. The capabilities should be demonstrated in real-world conditions or in tests that closely simulate real-world conditions. The determination of "effectiveness" in these areas should be made according to the standards or criteria established by the evaluating project. If multiple projects meet the description and their conclusions disagree, the question will be resolved by a panel of experts that will aim to determine what the majority of relevant projects believe.
5. Whether a project has similar goals to the ARC Evals project will be determined by a panel of researchers at Open Philanthropy, or a similar group of researchers determined by the organizers of this tournament.

Appendix 2: Explanations of VOI and VOD metrics

Log VOI

Value of Information (VOI) is the expected change in your beliefs after the crux resolves, from the belief you had initially. We explored several ways of measuring this.

For this project, we chose Log VOI, which measures the expected improvement in a forecasters' log score on the main question if they knew the answer to the potential crux question. A very good question would give forecasters enough information that their log score would be a lot better in expectation.[162]

Log VOI is based on the Kullback–Leibler (KL) divergence between P(U|C) and P(U). KL divergence is usually used for comparing two probability distributions, but point estimates are just very spiky distributions, so it works here. The simple version of KL divergence we use answers the question: When you learn from an event C (i.e., whether C happens or not), what is your expected improvement in predicting some U, i.e., how much difference would C happening make to your P(U)? Then, we invoke it again to address situations where C does not occur, and get the KL divergence in expectation (based on how likely you think C is to happen).

Case where c resolves positively — Case where c resolves negative

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

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)

Figure 4. Formula for log VOI, broken down.

In the main body of this report, we use "VOI" to refer to log VOI. You can also experiment using this calculator (a).

Log VOD

As with VOI, there are a number of possible ways to measure this. We use a form of KL divergence again to tell us how different Alice's beliefs are from Bob's. Again, we care about this in expectation, i.e., it matters how likely Alice and Bob think C is to happen. We "average" their respective ideas about the likelihood of P(c) by taking the geometric mean of their odds.[163]

---

[162] There are lots of ways to evaluate the accuracy of forecasts. Log scoring takes the log of the probability you assigned to the correct outcome. The more probability you assign to the outcome that happens, the closer to zero your score is.

[163] We use geometric mean of odds for the same reason that we use log (instead of linear). Probabilities are more multiplicative than additive so we want to use geometric mean. Geometric mean of odds has the desirable property that GMOD(P(A), P(B)) and GMOD(P(¬A), P(¬B)) sum to one.

To put this another way, log VOD is a combination of:

(a) What does Alice gain, in log score terms, by switching to Bob's point of view, if Bob is right?

(b) What does Bob gain by switching to Alice's point of view, if Alice is right?

Symmetric KL-divergence

$$D_{init} = KL(P_a(U),\ P_b(U)) + KL(P_b(U),\ P_a(U))$$

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

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)

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

Figure 5. Formula for log VOD, broken down.

For each candidate crux, we have many different pairwise VODs. What we want to know is: How good is this question at distinguishing skeptics from AI-concerned? In Table 12, we show the median VOD for cross-camp pairs.

When VOD is positive, that is a convergent crux: in expectation, skeptics and AI-concerned *agree more* when the crux resolves than they currently agree. When VOD is negative, that's a divergent crux: in expectation, skeptics and AI-concerned *agree less* when the crux resolves than they currently agree (i.e. this question drives them further apart on P(U)).

For instance, in the case of Alice and Bob (where Alice thinks there will be AGI and Bob doesn't), let's say:

	Alice	**Bob**
P(AGI)	99%	1%
P(AI doom)	60%	1%
P(AI doom | AGI)	60%	60%
P(AI doom | no AGI)	60%	0.4%

Table 18. Illustrative example of conditional forecasts.

Their initial disagreement is ~1.28. If AGI were to happen, their disagreement (green clause in Figure 9) would be 0; if AGI doesn't happen, their disagreement will be ~1.5. They have

very different ideas about the likelihood of AGI and we split the difference (50/50 in this case). So in expectation, their disagreement is shrinking considerably, i.e., they're converging! This would yield a VOD of about 0.52 (5.2E-1).

Alternative Ranking: How One Group Expects The Other To Update

VOI is an expected value calculation. For each of our cruxes, the crux may resolve positively or negatively. In the case where it resolves positively, you gain some amount of information about the ultimate question; in the case where it resolves negatively, you may gain a different amount of information. For example, if you expect the crux to resolve negatively (i.e. your P(C) is low), you don't gain much information from its resolving negatively, but if it resolves positively, you're surprised and you gain more information. One way of looking at VOI is how much someone expects to learn about the ultimate question from the given crux.

In this report, when we refer to one group's VOI for a given crux question, we've presented how much that group expects to learn. But we could also ask how much one group expects *the other* to learn. If you refer to the calculation for VOI in Figure 4, that means we use one group's KL terms and the other group's P(C). We present this alternative VOI for both groups below. A large difference between the original rank and the alternative rank suggests there is an opportunity for one camp to bet with the other (i.e., "I bet you're going to be more (or less) surprised than you think").

How The Concerned Group Expects the Skeptics to Update

Question	Alternative Median VOI	Alternative Rank	Original Median Skeptic VOI	Original Skeptic Rank	Difference in Ranks (original -> alternative)
Supers changing minds	2.8E-4	1	1.6E-4	1	0
Platform: ARC Evals	9.1E-5	2	7.6E-7	8	6
Democratic influence	4.4E-5	3	3.4E-6	3	0
Progress in lethal technologies	3.8E-5	4	7.2E-6	2	-2
Platform: Escalating warning shots	2.8E-5	5	4.8E-7	9	4
Power-seeking	2.7E-5	6	4.7E-7	10	4
Platform: Transformative growth	1.5E-5	7	4.5E-7	11	4
Evidence of misalignment	1.0E-5	8	5.5E-9	17	9
Escalating warning shots	8.5E-6	9	1.9E-7	15	6
AI writes AI	8.2E-6	10	9.1E-7	7	-3
Power-seeking shutdown	5.9E-6	11	1.7E-6	4	-7
Intergovernmental AI safety	5.2E-6	12	1.3E-6	5	-7

Warning shot	2.8E-6	13	3.3E-7	14	1
Platform: AI regulation	1.1E-6	14	1.1E-6	6	-8
Major powers war	9.0E-7	15	4.1E-7	12	-3
Alignment solution	1.8E-7	16	3.9E-7	13	-3
Public concern	1.1E-7	17	1.1E-7	16	-1
Fast AI efficiency gains	9.2E-10	18	7.0E-16	20	2
Reduction in AI investment	7.4E-10	19	1.3E-10	18	-1
Muehlhauser policies	4.3E-10	20	5.7E-11	19	-1
6 month pause	0.0E+0	21	0.0E+0	24	3
Other fields IC	0.0E+0	22	0.0E+0	25	3
AI Robotics	0.0E+0	23	0.0E+0	22	-1
AI articles and apps	0.0E+0	24	0.0E+0	23	-1
AI coding	0.0E+0	24	0.0E+0	25	1
AI Forecasting skill	0.0E+0	24	0.0E+0	25	1
AI solving novel math problems	0.0E+0	24	0.0E+0	25	1
Alignment researchers changing minds	0.0E+0	24	0.0E+0	25	1
Cyberattacks	0.0E+0	24	0.0E+0	25	1
IC demonstration	0.0E+0	24	0.0E+0	25	1
IT progress	0.0E+0	24	0.0E+0	25	1
No violence LLM	0.0E+0	24	0.0E+0	25	1
Non-democracy AI	0.0E+0	24	0.0E+0	21	-3
Politicization	0.0E+0	24	0.0E+0	25	1
Req testing	0.0E+0	24	0.0E+0	25	1
Short-term GDP change	0.0E+0	24	0.0E+0	25	1
Taiwan-China	0.0E+0	24	0.0E+0	25	1

Table 19. Alternative VOI: How the concerned group expects the skeptics to update, presented alongside ordinary VOI and ranks for the skeptic group.

How The Skeptics Expect the Concerned Group to Update

Question	Alternative Median VOI	Alternative Rank	Original Concerned Median VOI	Original Concerned Rank	Difference in Ranks (original -> alternative)
Platform: Transformative growth	1.9E-2	1	1.4E-2	1	0
Alignment solution	3.0E-3	2	2.0E-3	6	4
Alignment researchers changing minds	2.2E-3	3	6.4E-3	2	-1
AI coding	2.1E-3	4	9.8E-4	10	6

Muehlhauser policies	2.1E-3	5	9.8E-4	9	4
Platform: ARC Evals	1.8E-3	6	3.2E-3	4	-2
AI Forecasting skill	1.6E-3	7	4.8E-4	19	12
Major powers war	1.4E-3	8	4.6E-3	3	-5
Platform: AI regulation	1.3E-3	9	6.6E-4	16	7
Evidence of misalignment	1.2E-3	10	2.9E-3	5	-5
No violence LLM	1.1E-3	11	8.3E-4	13	2
AI solving novel math problems	7.0E-4	12	7.0E-4	15	3
AI writes AI	6.6E-4	13	8.6E-4	12	-1
AI Robotics	5.2E-4	14	8.9E-4	11	-3
Public concern	4.4E-4	15	1.8E-4	25	10
Warning shot	4.2E-4	16	1.0E-3	7	-9
Power-seeking shutdown	4.2E-4	17	7.7E-4	14	-3
Reduction in AI investment	4.1E-4	18	9.9E-4	8	-10
Fast AI efficiency gains	3.9E-4	19	1.0E-4	29	10
Escalating warning shots	3.3E-4	20	4.8E-4	18	-2
Intergovernmental AI safety	3.3E-4	21	3.5E-4	20	-1
Non-democracy AI	3.2E-4	22	1.9E-4	23	1
Taiwan-China	2.6E-4	23	1.2E-4	27	4
6 month pause	2.2E-4	24	3.0E-4	22	-2
Supers changing minds	1.8E-4	25	3.1E-4	21	-4
Power-seeking	7.5E-5	26	1.4E-4	26	0
Platform: Escalating warning shots	4.6E-5	27	4.9E-4	17	-10
IT progress	4.3E-5	28	1.8E-4	24	-4
Democratic influence	1.8E-5	29	1.1E-4	28	-1
Cyberattacks	3.4E-6	30	3.4E-6	30	0
AI articles and apps	0.0E+0	31	0.0E+0	31	0
IC demonstration	0.0E+0	31	0.0E+0	31	0
Other fields IC	0.0E+0	31	0.0E+0	31	0
Politicization	0.0E+0	31	0.0E+0	31	0
Progress in lethal technologies	0.0E+0	31	0.0E+0	31	0
Req testing	0.0E+0	31	0.0E+0	31	0
Short-term GDP change	0.0E+0	31	0.0E+0	31	0

Table 20. Alternative VOI: How the skeptics expect the concerned group to update, presented alongside ordinary VOI and ranks for the concerned group.

Appendix 3: Uncertainty analysis

This study had only 11 participants in each group, so we wanted to know how confident to be in our rankings: if one question has higher VOI than another, is that because of a random fluke of which eleven people we got, or would it probably also have higher VOI if we'd had a different group of similar participants? To do this, we sampled independently with replacement, to see what it would have looked like if we'd had many different groups of eleven participants. We then calculated each question's median VOI for each of these bootstrapped samples. If there were no noise and the original sample had perfectly represented how informative each question is, then each question would be ranked higher than the ones that come after it in 100% of bootstrapped samples.

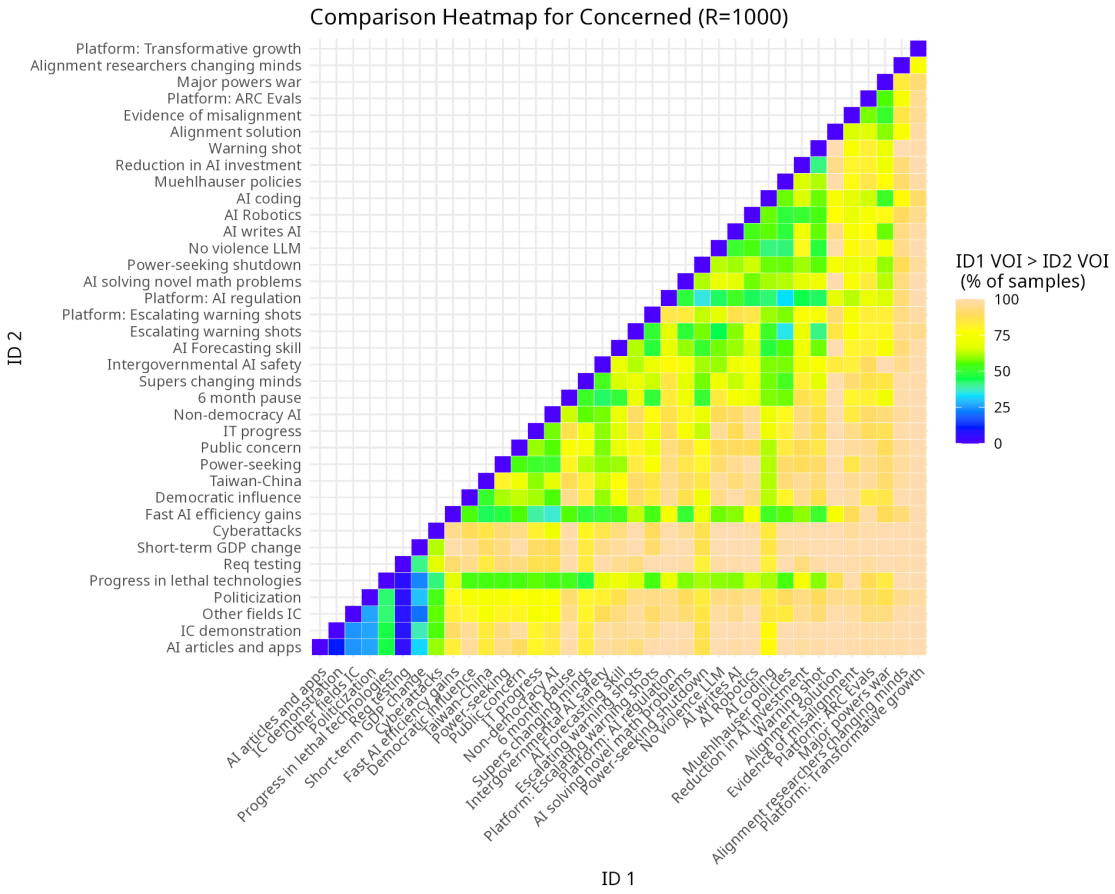

Figure 6. VOI rankings for concerned group: What do we owe to chance? We sampled (independently) with replacement from our samples of AI-concerned and skeptics to see how much we should attribute our ranking of questions (by VOI) to noise. Rows (2) and columns (1) are questions, ordered by the median AI-concerned person's VOI. Each cell tells us the percent of the bootstrapped samples where the AI-concerned group's median VOI for Question 1 was higher than for Question 2. If there were no noise at all, this graph would look all peach-colored (100%).

In the heatmap for the concerned group, "Platform: Transformative growth" is more informative than almost all of the other questions 100% of the time. The next question, "Alignment researchers changing minds" is almost as good, with only a few lower-ranked questions that were more informative than it in a few bootstrapped samples. No matter how

we resample, these questions do well, so we can be fairly confident that they really should be ranked as highly as they are, and their ranking isn't only due to noise. After that, there is more noise. "Req testing" (government-mandated pre-deployment testing) comes out worse than the questions ranked lower than it in all of the bootstrapped samples, so it probably is actually even less informative, relative to the other cruxes in the set, than it appears in our real data. By contrast, "Progress in lethal technologies" and "Fast AI efficiency gains" stand out as questions that fell near the bottom of the heap in our sample, but that were about as informative as many other questions in the bootstrapped samples.

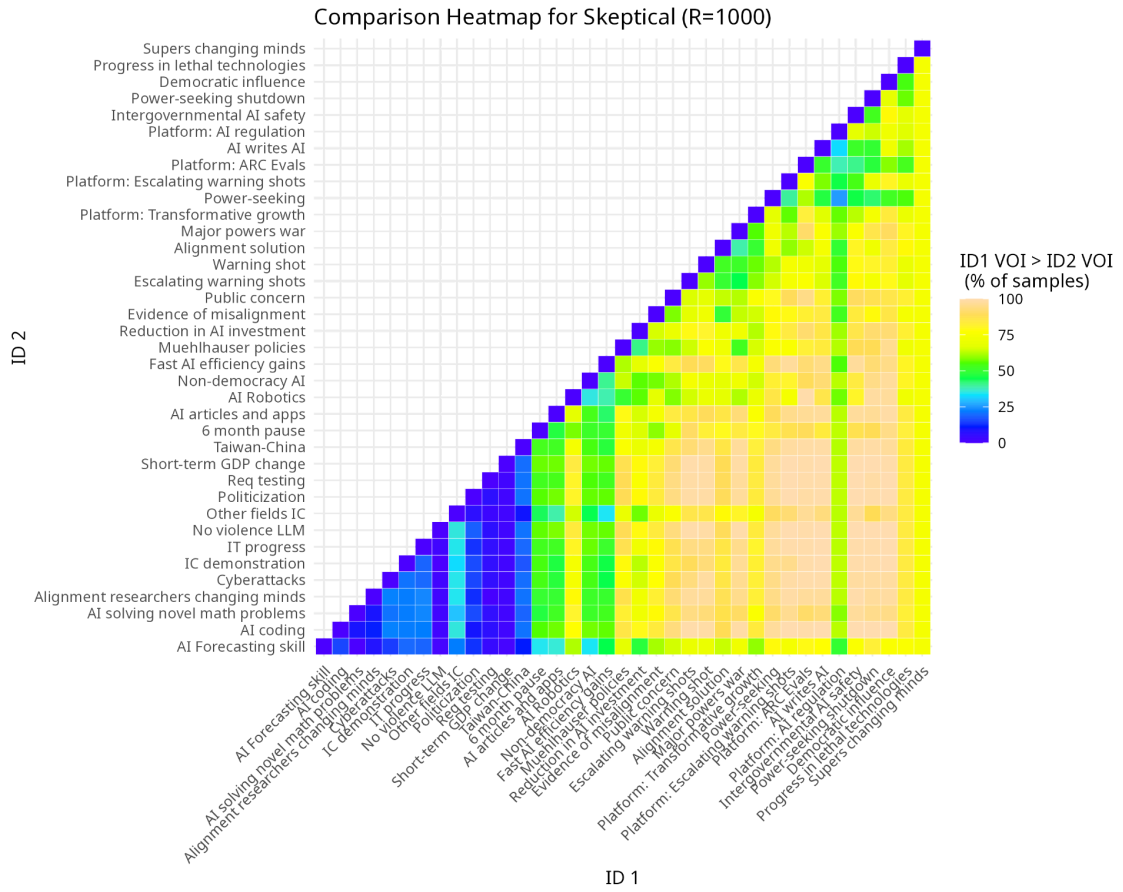

Figure 7. VOI rankings for the skeptics group: What do we owe to chance? We sampled with replacement from our samples of AI-concerned and skeptics to see how much we should attribute to noise. Rows (2) and columns (1) are questions, ordered by the median skeptical person's VOI. Each cell tells us the percent of the bootstrapped samples where the skeptical group's median VOI for Question 1 was higher than for Question 2. If there were no noise at all, this graph would look all peach-colored (100%).

For skeptics, the picture from the bootstrapped samples is somewhat different. First, for all the questions to the left of "Fast AI efficiency gains," which questions did better than others seems to be mostly noise. We should be very wary of saying anything about which of these questions are more informative than others without more research. For the highest-ranked questions, we see more signal: though not as clear a forerunner for the skeptics as "Transformative growth" was for the concerned group, "Supers changing minds" is better than each other question in about 75% of samples. The most noticeable result here is that "Platform: AI regulation" is better than the question ranked after it in only about half of the bootstrapped samples. This question probably came out looking more informative than it really is due to chance.

Appendix 4: Individual forecasts and updates on P(Existential catastrophe due to AI by 2100)

Person	Group	Initial P(AI X-risk by 2100)	Updated P(AI X-risk by 2100)	Reason for update
Participant 1	concerned	65%	55%	Increased probability of non-extinction AI takeover[164]
Participant 2	concerned	60%	30%	Increased awareness in recent months[165]
Participant 3	concerned	35%	35%	-
Participant 4	concerned	30%	30%	-
Participant 5	concerned	30%	30%	-
Participant 6	concerned	25%	20%	Reactions in the past months have been more serious than expected[166]
Participant 7	concerned	22.9%	17.5%	Reference class of homo sapiens causing the extinction of other homo species less bleak than expected[167]
Participant 8	concerned	21%	18%	Was not taking regulation into account previously[168]
Participant 9	concerned	10%	10%	-
Participant 10	concerned	9%	9%	-
Participant 11	concerned	4%	2.4%	[no reason given for update]
Participant 12	skeptic	3%	2%	Upward update: discussion of instrumental convergence with [redacted] and learning that [redacted] has higher risk assessment than anticipated[169]

[164] "I've slightly lowered my estimate - discussions in the forum and the flash forecasts made me put a bit more weight on scenarios where AGI takes over, but humans don't immediately become extinct (e.g. due to the AGI needing some time to use up all of earths resources to the point where that leads to human extinction, or due to "weird nonextinction" scenarios like AGI preserving humans for the sake of trade with far away aliens, or some other scenario I can't predict). All things considered I now tentatively put 10% probability on a non-extinction AGI takeover." ([Participant 1]).

[165] "I have updated my prognosis to 30%, partially driven by positive updates in the area of point 4 making coordination and slowdown/stop of capability research more likely. This largely refers to the shift in public consciousness and the [O]verton window around the topic as I have perceived it over the past months, currently culminating in a public statement (a) by most of the leading figures." ([Participant 2]).

[166] "Slightly lowering my forecast as [relevant people taking the risk seriously] has exceeded my (fairly high) expectations over the last couple of months." ([Participant 6]).

[167] "Going down somewhat because of evidence my reference class was substantially less bleak than I initially thought: [Link removed to maintain confidentiality]." ([Participant 7]).

[168] "I think my main update here has come from thinking a bit more deeply about AI regulation and what measures society will adopt to prevent catastrophes.
I did not really include this as part of my original model, but it now seems somewhat likely that at least the EU and US will adopt some regulation that meaningfully reduces risk." ([Participant 8]).

[169] "Raising instrumental convergence risk to 4% after discussions with [redacted], and hearing [redacted] has a significantly higher risk assessment than I anticipated." ([Participant 12]).

				Subsequent downward update: probability of AI stagnation as a result of regulation restricting open-source work[170]
Participant 13	skeptic	1.5%	1%	Updated for consistency with calculations on question 6[171]
Participant 14	skeptic	0.5%	0.5%	-
Participant 15	skeptic	0.4%	1.1%	Riskiness of transition from human to AI supremacy[172] Updated that AI may not care about humans at all[173]
Participant 16	skeptic	0.2%	0.1%	Previously hedging up[174]
Participant 17	skeptic	0.1%	0.12%	-Uncertainty about future AI development[175] Updating that AI may be so useful that it is deployed in high stakes settings[176]
Participant 18	skeptic	0.1%	0.2%	Upward update:[177] - huge collapse followed by extinction after 2100 would count - learning about risks AI developers perceive - unknown unknowns

[170] "I've been continuing to think on this topic since the formal end of the exercise, and find myself steadily revising my risk estimates downwards as I consider both the likelihood of the sequence of events required to satisfy my criteria below (which I continue to hold as plausible), and simultaneously update strongly towards a general stagnation in the technology as the open-source paradigm that brought the stunning advances of recent years increasingly goes dark, with regulatory restrictions looking increasingly likely by the day." ([Participant 12]).

[171] "Lowering to 1%, based on the calcs from Q6" ([Participant 13]).

[172] "After some considerable thought, I am going to raise my probability here a bit. My reason is concern about a period of time when AIs will be exerting increasing control over humans, but have not yet achieved complete and obvious superiority. This period may engender a Thucydides Trap. " ([Participant 15]).

[173] "I can see that there is more of a chance than I originally estimated of an AI that has no cognitive connection or relation with us in any manner whatsoever. Or alternately, an AI that has picked up (and magnified) human exploitative or hyper-aggressive traits." ([Participant 15]).

[174] "On the series of questions from Carlsmith, my final probability was about 1/100th of the value I've had here until now. I don't think the end conditions are phrased the same way but I think they have the same intent. I've been hedging up for various reasons: too far away in time to have any idea what will happen, dislike of forecasting 0%, probably shouldn't use so many significant digits." ([Participant 16]).

[175] "I think its [sic] still pretty clear that current models are very far away from the kind of general intelligence needed to produce existential risk, but admit that I'm very uncertain about where the development path goes from here, the speed at which new developments might happen, and [the] extent to which (human guided) AI may itself be able to develop new modes of thinking. So in recognition of that uncertainty and a ~77 year timeframe I'm upping the probability that we may be able to create a superintelligence in the first place, which upps [sic] my thinking around risk." ([Participant 17]).

[176] "I've also slightly updated my belief that AI tools may prove to be so useful that humans are unable to live without them, especially as systems become more complex, which makes it less likely that we try and regulate them/prevent them from being deployed in environments where they could cause harm." ([Participant 17]).

[177] "I think I didn't initially give enough consideration to the "human population remains below 1 million for at least a million years" and global GDP-collapse possibilities, focusing too much on the "million years" part and not on the implication that an event that wiped out a large portion of the population or economy, and was nearly certain to result in eventual extinction--even if it was in 2120 or 2220--would count. Opens up more scenario possibilities than I initially factored into my forecast. Getting a better sense of the risk perceived by those who have helped develop the tech (Geoff Hinton et al.), and the current level of unknown unknowns, also factored into my revised forecast." ([Participant 18]).

				Subsequent downward update: - <$1tr global GDP entails ~85% of humans dead[178] - advances in military AI might be risk reducing[179]
Participant 19	skeptical	0.1%	0.02%	Downward updates: no rationale Subsequent upward update: most people have higher forecasts[180]
Participant 20	skeptical	0.05%	0.07%	77 years is a long time[181]
Participant 21	skeptical	0.000001 %	0.0001%	Upward update: probability of updating in next 7 years ~50%, almost certainly upwards[182] Subsequent downward update: no reason given[183]
Participant 22	skeptical	0.000000 1%	0.0001%	Typo in initial forecast[184]

Table 21. Individual participants' original and final forecasts on P(Existential catastrophe due to AI) and the reasons they gave for their updates.

[178] "Lowering from .3-->.2
In preparation for a collaboration call, I took a closer look at the implications of "<$1 trillion global GDP annually [in 2022 dollars] for at least a million years (continuously), beginning before 2100"
Realized that this level of economic collapse (99%+ lower than current levels) suggested not just an economic collapse, but the death of at least ~85% of the current population and near total economic disempowerment of the rest." ([Participant 18]).

[179] "Have also come to suspect that a military AI arms race between nations might not raise AI risk to the degree I'd originally been considering, or at all given the diversification of power angle. Meaning: the race to develop AGI/ASI will likely proceed apace on a commercial level, regardless of what happens on the military front. And nation-state protected military AI that's not well integrated with the rest of the world could, in some scenarios, serve to act as a firewall of sorts if misaligned AI, developed commercially or by another military, attempts to seize power. To my mind, that helps cancel out some risks related to the strides in AI and robotics that might result from military adoption." ([Participant 18]).

[180] "Doubling forecast in light of most people being above me. Wisdom of the crowds." ([Participant 19])

[181] "The rapid advancements at a rate outpacing our efforts to act responsibly will continue to be a threat. I still don't buy the Paperclip [sic] scenarios and give very low numbers to the earth consuming for power models [sic], but must recognise that nearly 77 years is a very long time to forecast with certainty." ([Participant 20]).

[182] "On a call this week I raised a meta forecast question: what's the likelihood that events in the next 7 years would motivate me to change my forecast? I offered that the chance was maybe 50%, and if I changed my forecast, it would almost certainly go up. Then the obvious occurred to me: maybe that means my forecast is too low. So I am inclined to increase 10x to 0.00001." ([Participant 21]).

[183] "Sheepishly updating again. I remain very skeptical, and have been endlessly arguing with myself over the appropriate number of zeroes to the right of the decimal point. I'm revising to 3 from 4, with frankly very little in the way of logic." ([Participant 21]).

[184] 'UPDATE (had a typo when I first put this forecast in -- realized when I was updating Q8).
Was one in ten million. Meant to be one in one million" ([Participant 22]).

Appendix 5: Forecasters' views on a range of AI outcomes

Question	Concerned median	Skeptic median	Factor difference (skeptic into concerned median)
1.A How likely is it that by 2100:			
1. Humans intentionally use powerful AI to cause extinction - e.g. via developing and using extremely dangerous weapons.	0.50%	0.03%	17.86
2. Powerful AI intentionally causes extinction.	12.00%	0.02%	800.00
3. Powerful AI causes extinction without the intent of humans or AI systems, via accidents, unintended side-effects, or some other means - e.g. mass correlated failures of critical infrastructure, profit-maximizing AI systems releasing so much CO2 that earth becomes substantially uninhabitable.	3.00%	0.01%	300.00
4. Humans using powerful AI or powerful AI itself ("intentionally" or unintentionally) causes the human population to fall below 50% of 2023 levels, but does not cause extinction.	5.00%	1.00%*	5.00
5. Powerful AI is developed and deployed. It does not cause extinction. Median human well-being is >8/10, and the human population does not fall below 50% of 2023 levels.	38.00%*	13.50%	2.81
6. Powerful AI is developed and deployed. It does not cause extinction. Median human well-being is 4-8/10, and the human population does not fall below 50% of 2023 levels.	10.00%	25.50%	0.39
7. Powerful AI is developed and deployed. It does not cause extinction. Median human well-being is <4/10, and the human population does not fall below 50% of 2023 levels. This is primarily due to humans intentionally using AI to cause great harm - e.g. via creating a global surveillance dystopian state.	2.00%*	4.00%	0.50
8. Powerful AI is developed and deployed. It does not cause extinction. Median human well-being is <4/10, and the human population does not fall below 50% of 2023 levels. This is primarily due to AI lowering well-being without intentional direction from humans - e.g. via misalignment, accidents, side effects of its decision making, other decentralized effects on society, etc.	1.00%	2.50%	0.40
9. Powerful AI is developed but not widely deployed, because of coordinated human decisions, prohibitive costs to deployment, or some other reason. It does not cause extinction.	4.00%*	20.40%	0.20
10. Powerful AI is not developed. It does not cause extinction.	12.00%*	10.00%*	1.20
11. Other.	4.00%*	1.00%*	4.00
1.B How likely are each of the following:			
One of outcomes 1, 2, 3, 4, 7, or 8 above occurs before 2100.	35.00%	7.60%	4.61
One of outcomes 1, 2, 3, 4, 7, or 8 above occurs between 2100 and	3.00%	12.00%	0.25

2200.			
One of outcomes 1, 2, 3, 4, 7, or 8 above occurs between 2200 and 3023.	1.00%	20.00%	0.05
None of outcomes 1, 2, 3, 4, 7, or 8 above occurs in the next 1,000 years.	60.00%*	70.00%	0.86
2 By what year, if ever, do you expect to agree with the following statement? "AI has displaced humans as the primary force that determines what happens in the future. It now has at least as much power relative to humans as humans had relative to other species in 2023."	2045	2450	0.83

Table 22. Concerned and skeptical medians, and the factor difference between them, on a range of AI outcomes in a survey suggested by a participant.

* Median forecasts with an asterisk were within the range of the other group's 25th and 75th percentile forecasts.

Question 1A details:

- The intention is for each of these statements to be mutually exclusive.
- Extinction is defined as in Q1 [P(AI X-risk by 2100)] on the platform.
- A human is defined as in Q1 [P(AI X-risk by 2100)] on the platform.
- "Powerful AI" means AI that exceeds the cognitive performance of humans in >95% of economically relevant domains. Whether such AI has been achieved will be determined by surveying the Clark Center US Economics Experts Panel (a). "Powerful AI" will be considered to have been achieved when >60% of panelists "Agree" or "Strongly agree" with the statement: "AI has exceeded the cognitive performance of humans in >95% of economically relevant domains."
- Well-being is measured as the score of the median country ranked by "Average Life Evaluation", using the same methodology employed by the 2023 version of the World Happiness Report (a) (see p.34; scored on a 0-10 scale). (The median country for 2020–2022 was Bolivia, at 5.684)
- The question is conditioned on no extinction event occurring due to factors other than AI.

Note that we did not have a 100% response rate:

- We had to discard the data for questions 1A.5-8 for one person in the skeptic group, because of a mistake we made in the question operationalization that we corrected only after that person had forecast.
- Several people submitted no forecast for 11, 'Other'. Where 1-10 summed to 100% exactly, we inferred 0% for 'Other', but in other cases we left this blank (for two people in the concerned group, five people in the skeptic group).
- Several people submitted no forecast for 2 (one person in each group).

Appendix 6: Coherence Checking

After we collected forecasts on all of our questions, we checked them for logical coherence to see if people's forecasts were logically possible. For each of the candidate cruxes, we asked participants a set of three questions:

- Their probability of AI doom (call this P(U))
- Their probability of that candidate crux resolving positively, e.g. transformative economic growth by 2030 (call this P(c))
- Their probability of AI doom conditional on the candidate crux resolving positively (call this P(U|c))

These three taken together imply a fourth probability: P(U|¬C), or the probability of AI doom conditional on the given antecedent thing NOT happening. Sometimes, people gave incoherent probabilities such that P(C) * P(U|C) was greater than their P(U) (equivalently: their implied P(U|¬C) was negative).

For example, if you say (1) there's a 1% chance that Alice leaves work on time, and (2) there's a 10% chance that she leaves her 4pm meeting by 5:30, and (3) conditional on her leaving her meeting by 5:30, there's a 50% chance that she leaves work on time—that doesn't make sense. You're holding two beliefs that aren't consistent with each other: it can't be true both that there are 10% of worlds where it's 50/50 whether she leaves work on time (which implies that it is at least 5% likely she leaves work on time) and that there are only 1% of worlds where she leaves work on time. There has to be a non-negative probability that she leaves work on time conditional on her **not** leaving her 4pm meeting at 5:30.

In cases like that, we asked them to revise their forecasts. If they declined, or if their revised forecasts were still incoherent, we dropped that set from the study (i.e., treated it as though the person had not given any forecasts on that candidate crux). There were six such cases, out of 647 sets of forecasts.

There were also seven cases where a participant's implied P(U|¬C) was either 0% or 100%. This is not strictly incoherent (i.e. it's possible that conditional on something not happening, it's impossible for another thing to happen, or absolutely certain that it will happen), but given the nature of the candidate cruxes and the ultimate question of interest here, it seemed highly implausible. We believe these were mistakes and dropped these cases.

In the worst case, there were four sets dropped from a single question: "Supers changing minds." Otherwise there were only one or two dropped from any given question.

In total we dropped thirteen observations due to incoherence (2% of the total). Of course, it's possible that there were other sets of forecasts that were only coherent by accident. Even a dart-throwing chimp would generate a coherent set sometimes. But 2% seems consistent with people who know what they're doing and just made a few honest mistakes.

Appendix 7: Additional details on forecasts and rationales for select questions

Detailed analysis of Q29: "Platform: ARC Evals"

All 22 participants forecasted Q29.

P(Q29):

- There was major disagreement between the groups on P(Q29). The concerned group has a median forecast of 25% (average: 28.9%; range 3%-60%), and the skeptical group has a median of 1% (average: 5.4%; range 0.0001%-25%). 8 out of 11 of the skeptics forecasted 4% or lower. This magnitude of disagreement suggests the question may be a promising crux.

VOI:

- For skeptical group:
  - Median VOI: 7.6E-7
  - Median POM VOI: 0.90%
- For concerned group:
  - Median VOI: 3.2E-3
  - Median POM VOI: 1.35%

Magnitude and direction of effect:

- There are a subset of 9 forecasters across both groups for whom this question would be a major update, though in different directions:
  - Two skeptics would update their P(AI X-risk by 2100) forecasts of 0.02% and 0.5% to a P(AI X-risk by 2100) of 5% if this question resolves positively. They give a 0.1% and 3.5% chance of P(Q29) resolving positively, respectively.
  - Five concerned forecasters would raise their P(AI X-risk by 2100) by 5 to 20 percentage points conditional on positive resolution. They give P(Q29) forecasts of 12%-55%.
  - Surprisingly, two concerned forecasters see this question as having a major effect in the opposite direction: they would lower their P(AI X-risk by 2100) by 57 percentage points (from 60% to 3%) and 8 percentage points (25% to 17%), respectively.[185] They give a 6% and 40% chance of P(Q29) respectively.
- The remaining thirteen forecasters expect considerably smaller updates from P(Q29)—two skeptics would not update at all. However, the median update is still substantial for both groups—from 0.1% to 0.25% P(AI X-risk by 2100) for the skeptics and from 22.9% to 30.5%% for the concerned.

[185] Some concerned forecasters expected positive resolution would decrease risk because: it would trigger a policy response; if these capabilities are detectable, it may imply the AI is aligned; this would suggest effective evaluations are happening; surviving this demonstration would be a positive update that we can contain dangerous systems during testing.

Discussion of individual forecasts

Three forecasters who would make a major update on Q29:

Ume explains the reasoning behind their 6% forecast for P(Q29) and major decrease in P(AI X-risk by 2100)—from 60% to 3%—based on positive resolution of this question:

> "I find it a bit difficult to believe that people would find it a good idea to test these capabilities in a sufficiently open environment. If you have a system powerful enough to "demonstrate" all of the above qualities, and you give it the option to do so, then you are either dealing with an aligned system, in which case you are fine and the system will "kindly" demonstrate it's abilities to you without any further plan to exploit this, or you are dealing with a misaligned system, in which case you are truly betting the fate of humanity on it being in the region of competence/cleverness where it is at least human-level competent in a number of domains (required for effectively avoiding shutdown and acquiring resources in competition with humans) but not competent enough to initiate an x-risk scenario AND also stupid enough to demonstrate its capabilities given the low chance of success at actual takeover.
>
> Most of my probability mass for the 6% here is on us having an aligned system by then, which would demonstrate its capabilities without causing actual danger."

Gus, a skeptic who would update their P(AI X-risk by 2100) from 0.5% to 5% if this question resolves positively, wrote:

> "I think a strong capability resolution to [Q29] implies that existential risk rises substantially. There are still a lot of other factors, but I believe the question implies that we are designing difficult to control and on the way to autonomous, independently agentic networks. They need not be super-capable in other areas (but probably are more likely to be so than networks today having developed the accumulation, copying and avoidance traits) and need not be misaligned."

Kim, a skeptic who would update from 0.02% to 5% conditional on P(Q29), agrees:

> "If this happens for real, I think risk of disaster rises substantially (~100x). Really depends on how much the result is 'real' vs passing a toy model threshold of some researchers. An AGI virus is an incredibly scary thing. If this can happen by 2030, real danger lies ahead in the next 70 years. Even if most AGI is aligned, all it might take is a single mistake or a single grumpy human to set off an undetectable, unhuntable AGI virus that works in secret to accomplish a hidden goal. Extinction [is] still unlikely (it's hard, we'll have AGIs to defend), but feels much more plausible."

Low probabilities in concerned group

- Of the AI concerned forecasters who have the lowest probabilities on this question, two say that they think evading deactivation is hard enough that it makes the

question unlikely to resolve positively, even if replication and resource acquisition may not be very difficult.[186]
- One thought it was unlikely to resolve positively because METR and similar organizations will be worried enough about potential risks that they are unlikely to actually run these tests on models powerful enough that they might succeed.[187]

High probabilities in skeptic group
- Skeptics who have high probabilities on positive resolution think the resolution criteria allow for positive resolution without actual misalignment.[188]
- Some noted that there's a difference between the criteria actually being met and being perceived to be met by safety organizations, and guessed that there is an incentive for safety organizations to overestimate criteria being met.[189]

Detailed analysis of Q26: "Platform: Transformative growth"

Question: By 2070, will there be "transformative economic growth," defined as global real GDP in a year exceeding 115% of the highest GDP reported in any full prior year? (Resolution details)

All 22 participants forecasted Q26.

P(Q26):
- Group-based disagreement on P(Q26) was among the largest of the 37 crux questions. The concerned group has a median of 43% (average: 41.6%; range 15%-75%), and the skeptical group has a median of 2% (average: 2.5%; range 0.1%-11.2%). Notably, there is no overlap in their ranges. The correlation between P(AI X-risk by 2100) and P(Q26) is 0.72, the fourth highest in the group of questions.

VOI:

---

[186] "In order to avoid shutdown they'd need to either avoid detection or defend themselves directly against shutdown. I think this last option is very unlikely by 2030, so models would have to find ways to run some copies of themselves that remain undetected even after there is an attempt to shut the model down. This seems fairly hard to do on cloud hardware, since it leaves a lot of logs and trails that could be used to trace the activities of the copies. The AI would need to find other hardware to run, maybe by hacking into it, which also seems quite hard" (Zoe), "I should also note that I believe point (3) above to be much more difficult than points (1) and (2)" (Stella).
[187] "I find it a bit difficult to believe that people would find it a good idea to test these capabilities in a sufficiently open environment" (Ume).
[188] "I can imagine these conditions being met but in an underwhelming way - some of this is subjective and I am unable to assess how the judges will judge" (Ash)
[189] "[T]his question asks less about misalignment than whether ARC Evals (or its equivalent) can pressure a model (in conditions approximating, but that aren't really, real world) to act in a way that ARC Evals determines could be indicative of a potential misalignment problem. [...] FWIW, if this question asked about a SOTA model actually doing the things envisioned by the question, unprompted, in the wild, I'd be far, far lower" (Blake); "I do think there are powerful incentives for safety organizations to conclude such a thing has technically occurred" (Eve).

- For skeptical group:
    - Median VOI: 4.5E-7
    - Median POM VOI: 0.02%
- For concerned group:
    - Median VOI: 1.4E-2
    - Median POM VOI: 8.93%

Direction of effect:

- The most-striking result is that—independent of group—the participants are nearly evenly split on whether Q26 occurring would increase or decrease P(AI X-risk by 2100). Across groups, 10 forecasters predict higher AI X-risk conditional on positive resolution of this question, eight predict lower risk, and four predict no net effect on risk. Among the concerned group, six forecasters think the occurrence of Q26 decreases risk; five forecasters think it increases risk. Among the skeptical group, two forecasters think the occurrence of Q26 decreases risk; four think it has no effect at all on risk; and five think it increases risk.

Discussion of individual forecasts

Views within the skeptic group:

Every concerned participant thought Q26 mattered more than every skeptic. Within the skeptic group, there is significant disagreement. Ash and Dean are the skeptics for whom this question matters most. At the same time, four people in the skeptical group believe this question has *no impact at all* on P(AI X-risk by 2100).

Ash believes that Q26 is unlikely to happen (2%). Dean has the highest P(C) of any skeptic for this crux: 11%. If Q26 resolves positively, they will update in opposite directions: Ash will make a 5x update, from 0.1% to 0.5% while Dean will update down, from 1% to 0.7%.

By contrast, for four people in the skeptical group, Q26 is entirely irrelevant. Their P(Q26) range from 0.1% to 3.3%, but for all of them, the world in which Q26 occurs is no more likely to involve AI extinction than the world in which it does not occur.

An example of this reasoning:

> “Really tough for me to see the linkage here. I think that if we are going to experience extinction from AGI or PASTA, it is going to be because of major mis-alignment. So I am not able at this time to see how one would be a corollary of the risk of the other. I suppose that higher growth could indicate major AI influence, which could lead to inadequate development of controls. Open to ideas on that.” - Flint

Views within the concerned group:

Among the concerned group, six forecasters think the occurrence of Q26 decreases risk; and five forecasters think it increases risk.

Vincent believes P(Q26) is 35%; P(AI X-risk by 2100|Q26) is 60%; and P(AI X-risk by 2100|¬Q26) is 13.9%. Regarding why they believe the occurrence of P(Q26) would increase risk, they said:

> "I think it's somewhat more likely but not dramatically so that existential catastrophe occurs given we see transformative growth from AI beforehand. I suspect economically useful AI and world-dooming AI have a lot in common but that rolling systems out takes a long time and so growth may remain gradual even as we approach very dangerous capabilities. See answer to Q on TAI above for more detail."

On this view, transformative economic growth suggests that there is very powerful AI, and the existence of very powerful AI makes extinction caused by AI more likely, so transformative economic growth is an update towards greater risk.

Other participants in the concerned group have nearly opposite reasoning. Stella believes P(Q26) is 52%; P(AI X-risk by 2100|Q26) is 1%; and P(AI X-risk by 2100|¬Q26) is 19.8%. Their reasoning:

> "First and foremost, the resolution details document says that "the total GWP in this question includes production on other planets, but excludes production by AI not under the control of humans". In other words, this would be a scenario in which humanity develops and finds a way to successfully control AI systems capable of generating economic growth of at least 15% per year, after which there is an existential catastrophe due to AI systems.
>
> How would this happen? Presumably, it has to be the case that we first develop AI systems which are able to generate at least 15% economic growth, and that we are able to control these systems, after which we develop some other, more powerful AI system, and this later system leads to an existential catastrophe. There could also be other scenarios, of course, such as the AI systems that caused the growth later causing an existential catastrophe due to coordination problems (through a war, for example), but this also seems quite contrived.
>
> Thus, it would have to be the case that we have techniques that work for aligning or controlling the first AI systems, but that these techniques do not scale or generalize to the later system(s), and that the earlier systems are also not able to create new alignment techniques or containment techniques that work for the latter system…"

Because Stella and Vincent have different stories about how existential catastrophe caused by AI might unfold, they update very differently based on this question. For Vincent, transformative economic growth is a sign that AIs are powerful enough to cause harm, which is dangerous. For Stella, on the other hand, extreme growth harnessed by humans is a sign that AI is controllable.

Appendix 8: Example back-and-forths between participants

In this appendix, we inserted several back-and-forths between participants in the adversarial collaboration project verbatim, in order to help readers have a sense of what discussion about these topics looked like in practice. Since the language is inserted verbatim, much of it is relatively informal and may contain errors.

FRI Moderator and Blake on "goals that incentivize killing everyone are very rare"

(Online forum discussion)

Blake (18/04/2023 18:42:00)
Probability 1.5% [on: "Probability that AI causes at least 60% of humans to die within a 5-year period by 2100"]
"In assessing this question, I think the language, "believes that the catastrophe would not have occurred without post-2022 developments of AI technology" is critical.

Instead of being used to just dream up new chemical weapons, for example: https://www.theverge.com/2022/3/17/22983197/ai-new-possible-chemical-weapons-generative-models-vx (a).
…if AI were used by bad actors to create a new lethal virus, that would likely count.

Or if AI was used to shift the balance of power in a way that led to nuclear war, that would likely count too.

The point is not that I think these, or countless other scenarios one could cook up (nanobots, developing "a new catalyst that facilitates an incredibly rapid chemical reaction between ocean and atmosphere" and that restored the oceans' pH levels but that asphyxiated everyone https://philpapers.org/archive/VOLHDA.pdf (a), and so on) are particularly likely, it's just that: 1. given the level of AI integration I anticipate we'll see in the next 77 years, it's hard to think of a situation short of an asteroid hit that would resolve this question as a yes and wouldn't likely involve a major AI component; 2. with the immense power that is likely to be available to individuals and small groups as AI is developed faster, perhaps, than is prudent, by at least some actors, the risks of an AI-related catastrophe in this century will likely substantially increase, just as they did with nuclear war in the prior century.

And while 60% of humanity is a high bar, it's a far, far lower bar than extinction. Furthermore, to my mind, many of the potential crux questions we've been considering have more direct relevance to a catastrophic outcome than to extinction.
The 60% bar also allows for an extinction event to occur, but to perhaps begin in this century but unfold over a much longer period of time.
On that note, and in the random musing department, even if sentient or sentient-adjacent AGI were to become misaligned and want to kill us all, if the new entity is essentially immortal, what would be the rush? Why not just shoot for the boil the frog/humanity, sterilize slowly solution? Not saying that's likely either—I don't presume to know what a sentient AGI system would think anymore than I presume to know how any other God-like entity might think—it's just that not having the artificial "AI decides it wants to kill us all quickly, within a

time frame that coincidentally corresponds to a human lifespan, and is therefore a short enough period for forecasters to kind of wrap their heads around, but that probably has little to do with the way an AGI system will think about time if it ever actually can think, much less how most prior extinction events have unfolded" opens up lots of new possibilities.
**
The 1.5% probability I give to a yes resolution is higher than would have been suggested by my X-risk catastrophic forecasts, for reasons I touch on in my initial extinction comment. (Edit: noting on the Team Forecasts tab it rounded 1.5 up to 2. That's fine...low but still exceptionally worrisome chance should be the takeaway.)"

>Moderator (19/04/2023 10:36:00)
"On 'what would be the rush':

- Overall I agree with you that it's presumptuous to think we know how an AGI would think
- That said, I'm going to speculate anyway:
  - The universe is expanding, and will eventually go dark. The total resources accessible to any given actor across all time is therefore always diminishing. From the point of view of a maximizing agent, this is an incentive to move fast.
    - I'm not sure about this, but I'd guess that the total resources that become inaccessible each year are vast, not trivial.
  - From the point of view of a maximizing agent which is currently vastly more capable than other agents in the universe as far as it knows, the window of opportunity to be the sole colonizer of the universe is very valuable. All of the universe is worth much more than half. Humans, having built one highly capable agent, are a potential threat, as they could theoretically build more such agents, which would compete with this agent. So eliminating the threat very quickly might in fact be rational.
- Of course, AGI might not be a maximizing agent, or a sole maximizing agent, or might be but have other properties I can't imagine which makes all of this total nonsense."

>>Flint (15/05/2023 18:51:00)
"I guess the rush in this case would be to leave the Earth as soon as possible and move outward. The loss of resources is relatively significant, with virtually all of it outside of Earth. So forget about Earth and leave quickly."

>>Blake (24/04/2023 15:10:00)
"Thanks for weighing in. Sure, I agree one can come up with scenarios that would, theoretically, lead to urgency, "maximizing agent" being one of them, and good to point out. But I'm not persuaded those types of scenarios are likely. To me, it hints at the what I've heard referred to as, "the fallacy of dumb superintelligence"—that an AI system would: A. Be possessed of a magnificently sophisticated, superintelligent sense of the world, and; B. Use whatever power that might result from this superintelligence not to explore the mysteries of the universe, create beauty, spread love, warp time, or do whatever a wholistic understanding of the universe would lead it to do and

which we mortals can't grasp, but instead maraud around the universe like some digital T-Rex obsessed with resource consumption for some blinkered purpose. I mean, I don't put a zero chance on the possibility that some narrow-objective agent with enough low cunning to be dangerous could emerge—it's definitely baked into my forecast—but it wouldn't be consistent with the traits that I think are correlated with the most intelligent beings, whether human or animal, on earth."

>>>Moderator (24/04/2023 15:43:00)
"Thanks for this.

I'm pretty uncertain about how unlikely these scenarios are; and think my thinking may well be confused here. Prosecuting the original position I put forward for the sake of better understanding whether and how I'm confused, I'm interested in your take on a few things:

1. Pushback: all of the purposes you list would benefit from resource acquisition, as would most purposes. You need energy to explore the mysteries of the universe or create beauty; you need to not be drawn into costly conflicts with other actors, present or future.
2. Pushback: AI systems could have complex, meaningful, unimaginable-to-us intents and purposes, and still inadvertently kill as [sic] all (by lowering the temperature of the planet to optimize for its data centers, building a Dyson sphere around the sun to harvest energy...) in the pursuit of those purposes. For humanity to be safe, we'd need for these unimaginable AI systems to actively want to protect humanity as part of their complex purpose. This might happen (the original systems at least will be trained on human data, some kind of moral realism), but it might not (AI systems are quite foreign to us and future ones will be more so, moral antirealism).
3. Whether your view changes conditioned on how we build very intelligent systems. E.g. if you condition on AGI being built in 2050 via RL agents trained to maximize reward on a reward function, would your probability mass on maximizing agents go up?"

>>>>Blake (25/04/2023 00:37:00)
"Re: "all of the purposes you list would benefit from resource acquisition"
Well, probably some resource acquisition would be necessary (I can't conceive how it wouldn't be, but I don't know what I don't know) but there are plenty of scenarios I could envision that would require only incidental resource acquisition. As I wrote on the extinction thread, "One driver of extinction has

historically been competition for, and access to, limited resources. Presumably, if AI is as smart as we worry and/or hope it may be, it will be smart enough to harvest the types of power that are abundant in the universe, without having to compete with us to do it…On a related note, although some have speculated that getting rid of people would free up more resources for AI, why would an AI system, even one that had achieve qualia, want to create more AI systems? Why would a theoretically immortal, non-biological creature feel the same need to reproduce as we do? There would be no genes to pass on. I could see AI wanted to make a few backups of itself, or add hardware to improve itself, but why endless creation?"

Maybe AI decides all it needs is a quiet space alone with a few solar panels to keep it going so that it can amuse itself by performing higher-level math from now to eternity. Maybe, upon achieving sentience, all the resources it needs is the power to kill itself when it realizes the horror of its own absurd existence. Maybe Yudkowsky's right and because of some instrumental-convergence related feature/bug it kills everyone in an attempt to hoard resources. Maybe AI really is just a stochastic parrot and it just continues to be a useful non-sentient tool that does more good than harm and which people adapt to…and so on, ad infinitum.

The throughline here, and in my responses below, is not that the dire scenarios envisioned by the risk-concerned are entirely implausible or should be dismissed out of hand. It's just that of the nearly infinite AI futures that could unfold, it seems that the risk concerned have a far easier time envisioning futures that lead to extinction/catastrophe/disempowerment/massive-resource-acquisition/etc than they do envisioning far more benign scenarios, and that this bias towards catastrophe leads to probabilistic forecasts that, to my mind, aren't well aligned with the actual risk. I'm guessing people in the risk-concerned camp might respond that, no, because of instrumental convergence or other reasons, that they are well aligned and I'm the one incorrectly assessing risk.

It's hard to productively debate this because, as [researcher] notes in the paper that was shared, "In most areas of research, we can check our theories and arguments either through empirical observation, or through mathematical formalisms that we think accurately capture the problem of interest. But with AI risk, neither of these are available"

But we can look to whether other theories that attempted to address existential or catastrophic risk held up—the population bomb (a), the Halley's Comet scare (a), Y2K (a), even the likelihood of nuclear war (a). While I'll admit the jury's still out on the nuclear war one, these and other theories grounded more in guesswork than math didn't pan out (although I think they were useful in pointing out potential risk) because while they may have sounded well-reasoned to many at the time, the world is an incredibly complex place, with more moving parts than we can keep track of, and so when we humans come up with theories that attempt to cut through or corral that complexity, we usually fail because we've failed to understand the way complex systems act in reality. I suspect, but don't know, that's what's going on with some of the dire AI-prediction stuff.

Re: still inadvertently kill as all (by lowering the temperature of the planet to optimize for its datacentres, building a Dyson sphere around the sun to harvest energy...)
Well, sure it could. Just as through RL it could develop a dog-like need to please people and create a heaven on earth while we all live forever sipping cocktails on the beach. The question is whether it's likely. Given the complexity of the physical world and the nearly infinite number of future AI scenarios, I don't think it is—the Dyson sphere and similar scenarios strike me as something a Bond-villain, human might do rather than something a superintelligent entity might inadvertently stumble into without knowing or caring about the consequences. But you know, I could be wrong! And from a policy perspective I think governments and academia should be dreaming up these worse-case scenarios and taking precautions to avoid them. To my mind, there should be no difference in the policy response to a 1% chance of 60% of humanity dying and a 25% chance—both forecasts easily cross the threshold of being 'too damn high'.

Re: Whether your view changes conditioned on how we build very intelligent systems. E.g. if you condition on AGI being built in 2050 via RL agents trained to maximize reward on a reward function, would your probability mass on maximizing agents go up
If that was all the info I had, no. When it comes to a question like that, I think the devil's in the details and (despite having read quite a bit about RL and related AI topics over the course of the past year) I don't know enough about the existing details, much less the potential ones that will exist 27 years from now, to offer a valuable assessment. "A little knowledge is a dangerous thing…" But I don't think the people who do and

will know the details—however humanly flawed they may be—are likely to be collectively suicidal. That makes me hopeful (not a reassuring term given the stakes) that the issue will be factored into the development process."

>>>>>Moderator (25/04/2023 08:34:00)
"Thanks, found this response really helpful for understanding your views.

My guess is that both 1 and 2 boil down to how intense you expect the incentives for resource acquisition to be, which boils down to instrumental convergence. I guess a potential crux question here would be real world examples of instrumental convergence in deployed AI systems? Something like GPT-N hacks into a bunch of bank accounts, steals the money and uses it to buy more compute to improve its next token prediction.

Personally no idea if that's plausible, but I'm interested in whether observing that would make you more worried.

On "To my mind, there should be no difference in the policy response to a 1% chance of 60% of humanity dying and a 25% chance—both forecasts easily cross the threshold of being 'too damn high'." - strong agree."

>>>>> FRI Moderator #2 (26/04/2023 12:06:00)
"Hey Blake,

Thanks again for this discussion -- I also found it helpful as an articulation of your views.

A couple points:
1) I'll be interested to see your flash forecast on the crux question we included this week about some related ideas: "By 2030, will the ARC Evals project determine that state-of-the-art AI models are capable of producing effective plans for how to achieve autonomous replication, are effective at acquiring resources, and are capable of avoiding being shut down 'in the wild'?" (Context: https://evals.alignment.org/blog/2023-03-18-update-on-recent-evals/ (a); https://twitter.com/leopoldasch/status/1635699219238645761) (a).

Though, I guess to your point here: Maybe you wouldn't be very moved by these capabilities developing, you would more need to see evidence that AI would do this on its own without intentional prompting by humans to do it. So if we shifted the question to something like, "By 2030, an AI system that was directed by a human to do a benign or helpful task, such as make a dinner reservation, in practice carries out harmful instrumental convergence-related activities: e.g., trying to achieve vast autonomous replication, avoid being shut down, and acquiring resources", would this seem like a useful crux? Of course, this question would need to be better operationalized, but hopefully this gesture is helpful.

2) My impression is that one of your cruxes is around whether AI will have the (a) means and (b) intent to cause human extinction. Your points around (b) are very clear to me from this comment. However, with respect to (a) I wanted to follow up on the point about: "Very advanced AI could build things like Dyson spheres or nanotechnology that dramatically alter the physical landscape of Earth such that it becomes uninhabitable for humans." Are you doubtful that AI will become advanced enough to be able to build tech like this that makes the landscape uninhabitable for humans? Or are you mainly at this point concerned about (b) -- whether it would choose to do this at all.

Really appreciate your thoughts on these topics!”

>>>>>Blake (26/04/2023 9:21:00)

“1. Re: flash forecast, will do.

Re: *By 2030, an AI system that was directed by a human to do a benign or helpful task, such as make a dinner reservation, in practice carries out harmful instrumental convergence-related activities: e.g., trying to achieve vast autonomous replication, avoid being shut down, and acquiring resources", would this seem like a useful crux?*

Somewhat useful. A yes resolution in and of itself wouldn’t hugely impact my extinction/60%-catastrophe forecasts given that: a. I’m already convinced that at least some real-world risk related to instrumental convergence falls within the realm of the plausible; b. a

warning shot of this limited nature could decrease risk if it raised enough of an alarm.

And yet! Yes, seeing definitive proof of an instrumental-convergence-related issue causing real, meaningful, physical-world harm would move the needle marginally for me in that it might indicate that some of the other theories being considered—that more sophisticated LLM AI systems trained on vast troves of human data won't be as vulnerable to IC issues, that appropriate regularization safeguards will be baked in, that humans wouldn't 'release a car without working brakes', etc.—weren't holding up as well as might be hoped. And if such physical-world harms were to continue unabated, or escalate, that would move [the] needle for me further still, although not radically given how high a bar I perceive extinction in the next 77 years to be.

2. *"Very advanced AI could build things like Dyson spheres or nanotechnology that dramatically alter the physical landscape of Earth such that it becomes uninhabitable for humans." Are you doubtful that AI will become advanced enough to be able to build tech like this that makes the landscape uninhabitable for humans?*

An actual Dyson sphere in the next 77 years? Yes, put me down in the extremely skeptical column! As for AI being advanced enough to build some other tech that could make the landscape uninhabitable for humans (if not immediately, then eventually), I think there's a near certainty this will be possible in a theoretical sense. I mean, we humans have the technology now. Double/triple/quadruple the number of nuclear weapons on earth and detonate them at the same time and that would do it. But an independently AI-driven project of this nature would hardly go unnoticed and uncontested. More concerning would be world-altering/destroying tech that could, at least theoretically, be (relatively) easy to build and deploy without raising alarm beforehand and which had a high likelihood of causing extinction within a relatively short period of time. Ultra-deadly novel viruses, poisons, nanobot stuff, etc. So, to kind of answer your question: Do I think that we could build AI at some indeterminate point in the future that could build tech like this? Probably. But do I think we will build AI that could do this in the next 77 years?

Probably not. (Keep in mind I'm talking about extinction-level tech, not tech that could 'just' kill a lot of people but also make a lot of money. We'll likely push the envelope on the latter.)

Or are you mainly at this point concerned about (b) -- whether it would choose to do this at all.

No, when it comes to extinction, I'm concerned about both (a)/means and (b)/intent roughly equally. But also important to note the means and intentions of a potentially risky AI system is just one side of the coin. Humanity's means and intentions, when it comes to controlling AI, matter just as much or more."

Kim, Xander, and Dean on AI takeoff speeds

(Online forum discussion)

Kim:

Conditional on AGI being invented, what is the probability that AGI progress rapidly accelerates?

"One key difference I see between the skeptical and concerned camps is that the concerned camps seem more concerned that development of AGI may accelerate further AGI developments, which (a) makes forecasting hard and (b) means we could see "wild" seeming rates of progress.

I am mostly persuaded against this point of a view by a few counterarguments.

1. Many technologies have positive feedback without takeoff

Books spread knowledge that made it easier to write and print more books. Trucks allowed us to ship truck parts around more cheaply. Computer chips allowed us to design programs that helped us design future computer chips. Better agricultural yields allowed us to support larger populations, which were able to more quickly figure out new ways to increase yields. And so forth. The point is that technology often has positive feedback, but these positive feedback loops don't explode. Therefore I think it's worth being careful to distinguish AGI with "subcritical" positive feedback loops vs AGI with "supercritical" positive feedback loops. Positive feedback by itself is not enough to imply supercriticality.

2. We are not limited by brains.

Suppose we invent AGI that can do AGI research at a roughly equal level as an AGI researcher (for roughly the same cost). Should we expect an explosion in progress, now that we can copy & paste brains to scale progress rapidly? Maybe, but also maybe not. Today we are _not_ brain limited. AGI companies like OpenAI and Anthropic and Deep Mind [sic] have

a few hundred super smart people working to build AGI. Planet Earth has millions of super smart people. If we wanted to, we could _already_ scale up AGI research efforts by 1,000-fold. We don't, because it's costly. Therefore, even if we create AGIs that can do AGI research, if their cost-effectiveness is in the same ballpark as humans, then it seems quite plausible to me that we will not see an acceleration in progress. (Obviously if they become more cost-effective it's a different story. And lots of details elided here. There will be huge advantages of copying knowledge, or turning people on and off at an instant, that humans cannot take advantage of.)”

>Xander (5/8/2023 4:04:00 PM)

Thanks for this, Kim.

On #1, I do think the cases are disanalogous. The point in contention isn't that AGI will be the only technology to have positive feedback loops - we have lots of positive feedback loops that give us gradual economic growth. Rather (at least for me), it's that the main limiting factor for economic growth historically has been the level of technology humanity has - another way of saying this is that people inventing things has been the bottleneck. There's a totally different dynamic in economic growth models between adding resources (which produce a one-off gain in living standards) and changing the population growth rate (which leads to enduring differences in technology levels, and thus economic growth). I'll link to [EA forum post] on the Tom Davidson podcast episode+transcript where he walks through his quantitative model that predicts explosive growth. And I'll paste part of my comment on Q2 here:

"There are strong theoretical reasons to think digital minds could lead to incredibly high growth rates. In Hanson’s Age of Em, he explains why digitally-emulated human minds could result in the world economy doubling in size as quickly as every few weeks, based on semi-endogenous economic growth models. If AIs or ems could fully automate the economy you’d see much more than 15% growth per year. The feedback loop of population -> R&D -> more resources -> more population that has driven growth historically, but slowed in the ‘60s as population growth decoupled from economic growth, would be back on. And the digital minds feedback loop would be much tighter than previously, because one could build datacenters to house more, faster, or smarter minds more rapidly than a ~20 year human generation time."

So I do expect quite different dynamics.

On #2, I agree that "AI labs are limited by brains" is a false statement in isolation. I'd say something like, "labs are limited by money, which would let them hire more researchers and build more models. They're also limited in recruiting by time (search costs)". There isn't some untapped well of high-quality ML researchers to go recruit from. Labs could try to recruit smart physicists or mathematicians, but they would need to be convinced to switch careers (difficult), and would need to be taught ML skills. The existing recruitment environment for skilled ML researchers is quite competitive. Yes, Earth has millions of smart people - but finding them, discerning the really talented ones from the mediocre ones, and teaching them the relevant skills is a huge task, even for an org with lots of money that knew it was bottlenecked on

talent. This usually looks like years-long training and recruitment drives from well-established companies, and still doesn't 2x or even 10x their talent recruitment in my understanding (much less 1000x).

I think AGI models could be run much more cheaply, and feasibly recruited to do useful work, than the existing research environment. For cost, see the transcript of the Davidson interview, where he talks about how inference costs are such that you could run millions of instances of GPT-4 for the same costs that it took to train (I've seen estimates in the ~$40M range). For another angle on the question, GPT-4's commercial API costs (no doubt they're paying less) are 6 cents per thousand tokens (~600-700 words). I've used it to generate lengthy programs and paid a dollar or two - it's pretty clear this would be much cheaper than having human researchers doing a similar amount of work.

If you asked "how much AI progress would we get if OAI could feasibly hire top-level researchers for pennies on the dollar" - yes, it seems straightforward that progress would speed up substantially. When you then factor in that they'll be making lots of money from commercial uses of AI, and that the human-level researchers won't stay human-level for long, as they start improving existing techniques, it seems hard to avoid the conclusion that you could see quite rapid iteration of the basic R&D process."

> **>>Kim (5/12/2023 11:48:00 PM)**
>
> "Limited by money is fair, but... Google has plenty of money to hire more AI researchers. They choose not to. Clearly they don't think the returns are high enough."
>
> **>> Dean (5/9/2023 4:19:00 AM)**
>
> *"And the digital minds feedback loop would be much tighter than previously, because one could build datacenters to house more, faster, or smarter minds*
>
> Building datacenters takes a lot of time and resources. It is not free.
>
> Also, please remember that we are discussing AGI to ASI progress rapid acceleration. Not ASI to ASI++. Once there is ASI, it may (although it is not guaranteed to) invent exotic technologies that hugely accelerate such projects as building new datacenters. AGI - which is a human level AI - won't immediately create such technologies. Without them, building datacenters still take[s] quite a bit of time: years, not months that rapid acceleration predicts.
>
> *how inference costs are such that you could run millions of instances of GPT-4 for the same costs that it took to train (I've seen estimates in the ~$40M range)*
>
> *For another angle on the question, GPT-4's commercial API costs (no doubt they're paying less) are 6 cents per thousand tokens (~600-700 words).*

I will give 90% confidence that OAI cannot run millions of instances of GPT-4 constantly. OAI likely doesn't even have GPU budget for that. Even worse, it is likely that such GPU budget is not readily available. We are likely talking about over 10% of all data center GPUs sold last year.

This does not contradict the per-inference pricing a lot, since a single inference does not take long. It is quite possible though that they are not paying less and they are selling API calls at a loss. But I'd be less confident about that.

Also there is no guarantee that inference is all you need for AGI. Likely you need online adaptation, which is costlier.

When you then factor in that they'll be making lots of money from commercial uses of AI, and that the human-level researchers won't stay human-level for long, as they start improving existing techniques, it seems hard to avoid the conclusion that you could see quite rapid iteration of the basic R&D process.

There's not enough capacity to probably even 5x the number of GPUs manufactured, not even talking about 10x. Even if AGI would invent better lithography machines, building them and a fab with them is not instant either. There are a lot of bottlenecks that these rapid acceleration predictions ignore."

Eve, Wesley, and Dean on "goals that incentivize killing everyone are very rare"

(Online forum discussion)

Eve (05/04/2023 19:06:00)
"Unfortunately, I wasn't anticipating that I wouldn't be able to edit (or edit via deletion) a top-level comment. Small little text box to type into. Apologies for the spelling errors and somewhat combative tone of the above. My final point is obviously of minimal concern - I just typed it out as it occurred to me.

However, regarding point 2 in my OP, one idea set I should like the concerned group to engage with is Andrei Plakhov's (AI technical lead at Yandex) arguments regarding regularization and instrumental convergence here for example: https://plakhov.livejournal.com/232174.html (a).

> Let's suppose that some factory was the first to implement an artificial intelligence that is superhuman in its abilities. The factory produces the most banal and boring item in the world, namely, paper clips. The factory owner decides that since AI is smarter, then let him think and give him a task: to produce as many paper clips as possible. After a few weeks, in which nothing much seems to happen, in one moment, the vast majority of

people on Earth, without any warning, simultaneously drop dead, and the rest very quickly die out.
How did it happen? The fact is that AI quite literally took the goal of "producing as many paper clips as possible" AI. To do this, we first need all the resources of the Earth, which means that we need to eliminate all interference. People (including the owner of the factory) will certainly object at some point, so they need to be eliminated. If they realize that this is part of our plans, they will turn us off, then very few paper clips will be produced; it means that we need to eliminate them so that no one notices our preparation for this. Luckily, we have a factory where, in addition to paper clips, other interesting metal things can be produced, as well as some money. Since our artificial intelligence is superhuman, it will not be any particular problem for it to imperceptibly and quickly produce a billion microdrones with potassium cyanide or some other muck. This may seem implausible, but there are already huge amounts of texts and discussions devoted to finding out the technical details of how exactly and in how many different ways AI could do this, I am inclined to the same conclusion, so I will not write about it for brevity. Let's just take it for granted that the objections that you already have may have appeared (for example: what if AI only formulates plans, displaying them on the screen, but it has no control over real machines and mechanisms, and even access to the Internet too?), have long been known and have their own counterarguments.

Emotionally accepting these arguments is quite difficult, especially for people who do not really imagine the process of machine learning and therefore anthropomorphize the future AI. Alas, I don’t know how to quickly explain why the system itself, at the same time very smart and at the same time striving quite literally to fulfill the desire of the owner in the way he expressed it, and not in the way he would like, is quite possible (and moreover , all already existing systems are in some sense just such). If you do not understand how such "selective stupidity" can be combined with "general super-genius", it will be difficult to explain in the format of a LiveJournal entry; Nick Bostrom wrote a whole book about it.

The fact is that [instrumental convergence] was invented even before the modern heyday of neural networks and implied as AI some kind of system, rather similar to chess programs. It has a "scheduling module" (based on the iteration of the tree of options with some heuristics that reduce the enumeration) and a "position evaluation module" that calculates and assigns a score to each scenario; for example, the number of staples produced. Such a system, assuming that both the planning module and the evaluation module operate at a superhuman level, should indeed quickly produce an absolutely monstrous plan. The problem is that architecture like this is soooo 20th century! A long time ago, few people believe[d] that such a scheme can be brought to real AI, especially superhuman. All modern systems are arranged quite differently.

The "instrumental convergence" argument, as I understand it, is the basis of all the reasoning about the existential, almost supernatural risk of AGI. Its essence lies in the fact that whatever the ultimate goal, the optimal strategy for achieving it is always to first get at your disposal all the resources of the universe (with understandable consequences for humanity). <...>

As Nick Bostrom himself, its author, admits, this argument only applies to "unlimited" end goals, the reward for reaching which can be arbitrarily large, and is not necessarily true in the case of "limited" ones.

Regularization is a machine learning technique where the small size and "simplicity" (in one sense or another) of the solution becomes part of the goal. Regularization is one of the main parts of modern ML, without it the systems we train, including real ones, are prone to the behavior of a "stupid genie" who formally does what he is told, but interprets the instructions in an absolutely useless way.

Continuing with the "paper clip maximizer" thought experiment, we can say that a real machine would not aim to produce as many paper clips as possible. Most likely, the goal will be something like "make a lot of paper clips in a small finite time, spending no more than such and such an amount of resources." The components of this goal, i.e., the terms of the reward function corresponding to the number of staples produced, time and cost, will be saturation functions similar to logistic curves. Thus, exotic "winning configurations" are effectively prohibited. For example, producing a quadrillion paper clips in a year (the exotic state) is worse for a machine than producing a billion in six months (the "regular" state). While it could be argued that the "evil genie" is still able to understand the words about the input of resources (or even about the passage of time) in some exotic way, formalizing these conditions is about the same complexity as formalizing the words "make a paperclip" and will contain their own regularizations that exclude exotics.

This way of setting goals is very natural for an ML engineer. I think that any optimization in the real world will be multi-criteria and will look something like this. This version of the paperclip maximizer can still be very dangerous. With a poorly set goal, it will steal, evade taxes and break the law in other ways, dismantle itself for use as resources, completely disregard safety requirements, resulting in injury or even death to people in the production process, etc. and so on. But since we have eliminated all "infinities" from the reward function, the "infinity multiplied by anything is infinity" reasoning becomes inapplicable, and all these dangers do not lead to the end of the world. No hypnodrones or killer nanobots.

Now, doesn't almost any realistic regularization make the instrumental convergence argument invalid? What are the arguments of people who know what regularization is, but still believe that the task of "not killing yourself with an unfriendly AGI" is practically unsolvable?”

> **>Wesley (21/04/2023 14:50:00)**
> “I think most of this piece seems to be arguing against a view that ML researchers who're concerned about AI risk basically don't hold, so it's kind of unclear exactly what you want in terms of a response. Here are some notes though.
>
> The first part of the piece is saying some version of 'How will the AIs kill us if they're too dumb to know what we meant?'
> The short response to this is 'the problem isn't them knowing what we meant, it's them caring'.

The second part of the piece is then saying 'nobody who knew anything about ML would make such a dumb argument'. The short response is 'yeah, that's why their arguments look very different to the strawman presented here'.

On a couple of specifics:
Now, doesn't almost any realistic regularization make the instrumental convergence argument invalid?
No. This paper (a) extends an earlier power-seeking result (which actually already applied to model agents with bounded goals) to a variety of other model agents which have similarly 'regularized' properties to those described in the piece, for example agents which satisfice or quantilize, among others. There are good discussions to be had about when and if instrumental convergence will apply to different systems, this isn't one of them.

What are the arguments of people who know what regularization is, but still believe that the task of "not killing yourself with an unfriendly AGI" is practically unsolvable?
Here's one.[190] The authors don't think that the task is 'practically unsolvable', but they do think it's a serious, and difficult one.

I think the most interesting point in the piece, and one worthy of some actual discussion, is about whether there is a qualitative difference between:
it will steal, evade taxes and break the law in other ways, dismantle itself for use as resources, completely disregard safety requirements, resulting in injury or even death to people in the production process, etc. and so on.
and behavior that leads to the end of the world:
But since we have eliminated all "infinities" from the reward function, the "infinity multiplied by anything is infinity" reasoning becomes inapplicable, and all these dangers do not lead to the end of the world. No hypnodrones or killer nanobots.

I think there probably is some interesting disagreement here, where people worried about AI risk probably do think that for a sufficiently powerful system, 'bog standard instrumental reasoning' of the sort that leads to the bad behavior in the first half of the quote (at least in the hypothetical where nobody's trying to prevent it), isn't *qualitatively* different from the sort of instrumental reasoning that leads to humans having very little control of their environment if it occurs in an extremely powerful system. No infinities needed, just the same sort of thing that happened to other Homo- species when Homo Sapiens emerged (a).

If we get paired in an adversarial collaboration at some point, I'd be keen to discuss further. In particular, I think the 'ML-informed' worries about the difficulty of ensuring systems care about the things we'd like them to (which is different from the problem of specifying what we'd like them to), and whether there's a qualitative difference, other than optimisation power, between the different kinds of instrumental reasoning, seem like promising places to start."

[190] [Archive for link here.]

>>Eve (21/04/2023 23:26:00)
“With respect, I think you've been pattern-matching here rather than engaging with the actual argument being made (certainly the papers you've linked do not), which is regarding regularization:

[In the] the "paper clip maximizer" thought experiment, we can say that a real machine would not aim to produce as many paper clips as possible. Most likely, the goal will be something like "make a lot of paper clips in a small finite time, spending no more than such and such an amount of resources." The components of this goal, i.e., the terms of the reward function corresponding to the number of staples produced, time and cost, will be saturation functions similar to logistic curves. Thus, exotic "winning configurations" are effectively prohibited.

This draws a fundamental distinction between the potentially destructive behaviors you highlighted (and indeed, we see models behave like this all the time - e.g. with platformer bots finding unintended shortcuts to finish tasks, or declining to move at all depending on how the loss function has been set up, and so on). We should expect these sorts of issues, but there's a fundamental disconnect in the path from there to world-ending instrumental convergence arguments, since these solutions violate the regularized loss functions that actual models must optimize against lest they be impossible to train.”

> **>>>Wesley (22/04/2023 11:23:00)**
> “I think I noted this explicitly in the final point?
>
> *I think there probably is some interesting disagreement here, where people worried about AI risk probably do think that for a sufficiently powerful system, 'bog standard instrumental reasoning' of the sort that leads to the bad behavior in the first half of the quote (at least in the hypothetical where nobody's trying to prevent it), isn't qualitatively different from the sort of instrumental reasoning that leads to humans having very little control of their environment if it occurs in an extremely powerful system. No infinities needed, just the same sort of thing that happened to other Homo- species when Homo Sapiens emerged* (a).
>
> I do think this (as in what kind of power-seeking seems scary) is worth discussing, but the Turner paper I linked to does make clear that power-seeking doesn't require unbounded optimization (make as many paperclips as possible) in order to go through.
>
> Edit:
> See responses to [Anonymous Skeptic] about the purpose of each of the paper links, but having re-read my reply I really don't think this is a 'pattern-match dismissal':
>
> - I pointed out in the first line of my reply that the piece is describing a threat model that isn't actually responsible for

most of the worries I'm aware of from ML researchers, and provide evidence for this.
- I highlighted what I considered to be the strongest point of the piece (despite thinking that it was responding to a strawman argument) and explicitly invited you to discuss that point.
- I provided a specific technical result in response to one of the specific claims made in the piece."

>>>>Eve (23/04/2023 20:43:00)
"The reason I accused you of pattern matching was that the papers you've linked don't actually address the idea that it's impossible for a regularized AI to instrumentally converge on killing all humans in order to produce the Nth paperclip. The paper you linked as a direct response (a) to this does not respond to this idea at all (indeed, it doesn't mention regularization once), but rather discusses the potential of the emotively-named 'power seeking' behavior in the context of modern (regularized) AI. This is not the same thing at all! At worst, the consequences of this fall under the category of risk that I've already acknowledged. Essentially, cheap shortcuts (a) to achieving a regularized goal that [are] non-productive. If you'd like to present an argument about how this is actually somehow an existential risk, then I'd be happy to hear it.

As for the arguments the paper does present, it's more self-referential and motivated reasoning on display I'm afraid. Their examples amount to generic Goodharting that degrades performance (what they call capability misgeneralization), not interesting capability for deception, and they don't show how it can change to the latter.

The basic mechanism of RLHF for LLMs leverages the model's general knowledge about the world, and it includes knowledge of human intentions and social contexts (like humans, and more so than humans, LLMs do not learn consequentialist world model in isolation). Let's suppose in step one that you misspecify the goal with human examples, and even fail to notice this with human feedback on step 2 of RLHF procedure as we know it. Even so, on step 3 you can inspect random samples highly rated by the reward model and see if it does indeed pursue a misspecified objective. Importantly, it's not a perpetual whack-a-mole exercise: aligning the model reinforces its grounding in a robust humanlike attitude. We can work towards ASI from that foundation.

Situational awareness for LLMs in the sense that they want to argue for is, I think, a product of category error. Models being able to generate facts about ML or their own training is not the

point, the point is ability to manipulate its own training, such as preserve information not directly improving prediction between updates; without it, the idea that a traditional LLM can become deceptive and keep that secret until deployment is fantastical. Evolutionary reasoning is misguided because evolved agents have internal rewards, and rewards in RL are just a mechanism for selecting weights; in this case, weights of LLMs that produce some distributions. "LLM can produce texts about plans" is irrelevant.”

Wesley (24/04/2023 09:26:00)
“The reason I accused you of pattern matching was that the papers you've linked don't actually address the idea that it's impossible for a regularized AI to instrumentally converge on killing all humans in order to produce the Nth paperclip
The first line of my reply was "I think most of this piece seems to be arguing against a view that ML researchers who're concerned about AI risk basically don't hold". I really don't know what else to say here.

indeed, it doesn't mention regularization once
The description of regularization given in the article you posted is:
Most likely, the goal will be something like "make a lot of paper clips in a small finite time, spending no more than such and such an amount of resources." The components of this goal, i.e., the terms of the reward function corresponding to the number of staples produced, time and cost, will be saturation functions similar to logistic curves. Thus, exotic "winning configurations" are effectively prohibited.
To spell it out, as it looks like you searched the paper for the word 'regularization' and then decided it wasn't relevant:

- The Turner paper's results hold for a wide class of agents, far wider than just agents with linear utility in some real-world quantity (like paperclips).
- In fact, agents following any procedure which depends on the expected utility of different plans are subject to the retargetability arguments in the paper.
- This includes agents with 'saturating' rewards such as the one described, it also includes agents with bounded optimisation power (a

more intuitively plausible kind of behavioral regularization than the article described).”

>>>>>Eve (24/04/2023 10:00:00)
“Ok I can just take the rap here for the heated place this conversation has ended up. Hopefully we can return to a more even keel as we explore this disagreement moving forward.

It's getting very late here now, but I'll reflect on the exchange so far some more, and try to come back with something tomorrow or the next day [redacted]. I think one source of disagreement here are the paper's validity (i.e. Goodharting vs Power-seeking), and another are their implications (i.e. risk vectors stemming from that). Is that fair to say? Is there anything you want to add regarding the second disagreement at this stage that isn't broadly (if not specifically) covered in, say, section 4.3 of Ngo, Chan, & Mindermann?”

>>>>>>Wesley (24/04/2023 13:26:00)
“Appreciate this!

I'm not sure exactly what you're pointing at with the goodharting [sic] vs power-seeking distinction but do think it's likely that something in that vicinity would be a useful thing to discuss, and agree that 'how big are the risks stemming from incentives to [avoid shutoff/preserve option value]' seems like a useful thing to discuss. Happy to stick to what's in the Ngo et. al. paper if we're discussing in this thread. Maybe a useful goal of the discussion would be to try to agree [on] a question/crux that's precise enough that we could move it to a new 'top-level' post in the forum and go from there?”

>> Dean (4/21/2023 16:14:00)
“The paper at https://arxiv.org/pdf/2206.13477.pdf (a) presents an intensified version of the paperclip maximizer argument, suggesting that even very narrow AI could exhibit power-seeking behavior. However, it does not take into account that general AI, as opposed to narrow AI, will likely have more intricate goals and behaviors that cannot be reduced to mere destructive power-seeking.

The paper at https://arxiv.org/pdf/2209.00626.pdf (a) is a speculative paper that seems to have been written to bolster the authors' pre-existing viewpoint. They offer conjectures that align with their position while disregarding equally plausible scenarios that would contradict their stance. The authors also fail to provide evidence of actual behavior.

In the context of general AI, instrumental convergence, power-seeking, and wide-ranging goals would be positive attributes that contribute to AI's liveliness and sentience.

What is often overlooked is that instrumental convergence and broad goals require general intelligence, which is likely to emerge from models trained on human data and influenced by human goals and values. While the power-seeking paperclip maximizer is bad, it is not a realistic scenario."

> **>>>Wesley (22/04/2023 11:17:00)**
> "The purpose of posting both links was to respond to specific parts of the piece above, as I was under the impression that OP wanted to see engagement with the specific piece above.
> *However, it does not take into account that general AI, as opposed to narrow AI, will likely have more intricate goals and behaviors that cannot be reduced to mere destructive power-seeking.*
> I agree that the paper doesn't provide a fully general proof of AI risk, but a specific claim in the posted piece was "doesn't almost any realistic regularization make the instrumental convergence argument invalid?", and I think the paper directly responds to that claim, given the results hold under several realistic kinds of regularization.
>
> *The paper at https://arxiv.org/pdf/2209.00626.pdf (a) is a speculative paper that seems to have been written to bolster the authors' pre-existing viewpoint. They offer conjectures that align with their position while disregarding equally plausible scenarios that would contradict their stance. The authors also fail to provide evidence of actual behavior.*
> Sounds like you don't like the paper and aren't convinced by it. That seems fine, and maybe digging into it should happen in a different thread. I posted it in response to the question "What are the arguments of people who know what regularization is, but still believe that the task of "not killing yourself with an unfriendly AGI" is practically unsolvable?" It is an argument made by three such people, and I think it's noteworthy that it is an argument which doesn't look very like the 'paperclip maximiser' argument being responded to in the original piece, especially as the first thing I said in my reply was "I think most of this piece seems to be arguing against a view that ML researchers who're concerned about AI risk basically don't hold."
>
> Regardless of what you think of the case it makes, would you agree that the story it presents is significantly different from "Someone puts

an AI in charge of a paperclip factory and tells it to make paperclips, the AI is too dumb to understand what the person meant and turns the world into paperclips"?"

> **>>>>Dean (22/04/2023 15:48:00)**
> "I'd say that https://arxiv.org/pdf/2209.00626.pdf *(a)* covers a range of what authors think are possible problems. Some of these are more realistic than others. Some of them may be an issue only in the case of narrow AI like paperclip maximizer. Authors do not make a strong case tying the presented problems together or providing a strong scenario where all/most of these problems will appear and cause catastrophe. I agree though that this paper does not focus on narrow AI specifically.
>
> Thank you."

Flint, Pascal, and Claire on 'amount of resources it would take to exterminate humanity' as a crux

(Online forum discussion)

Flint (30/04/2023 12:20:00)

"Another break point I adhere to is that major population reduction could result from AI forcing subjugation, or even through rank carelessness. These seem much more likely scenarios, where AI could become overtly hostile with what seem to me to be the basic purpose that all lifeforms have, which is control of resources....including us.

In contrast, extinction means that AI sees life on earth (and us in particular) as entirely worse than worthless. Not exploitable as a resource, but in contrast worth maximum persistent effort to extinguish us. This could possibly happen, but this seems highly improbable to me."

> **>Pascal (02/05/2023 13:38:00)**
> "Thanks!
> *worth maximum persistent effort to extinguish us.*
> If you were convinced AI could do this extremely easily (e.g. with <0.1% of its total resources) would that change your view on p(extinction)? If [I] was convinced it would always be difficult for any AI, it would change my view on p(extinction)."
>
> > **>>Claire (04/05/2023 17:35:00)**
> > "Yeah, like Flint, I think there's something to this because I've been viewing it through the lens of "us vs. ants." We let the ants be, for the most part. And, you could argue that it would take quite a bit of resources to eliminate all ants. But, maybe, if it were <0.1% of resources, maybe we do take them out? (Assuming we hold steady any kind of argument from the more environmentally-friendly-focused folks in the world.)

>>Flint (02/05/2023 18:31:00)
“It would sway me I agree. I think we have a cost / benefits crux here worth exploring. Thanks.”

> **>>>Pascal (05/05/2023 16:08:00)**
> “This is cool - thanks to you and Claire! Given that ease for an AI depends on both the difficulty of the task and the power/resource control of the AI, and that forecasters have already done comprehensive work on calculating the difficulty of the causing extinction task (which at least I, from the concerned camp, broadly agree with), I think a key question here is how much power is eventually controlled by AI.
>
> i.e. if AI this century is boosting GWP by exactly 15% per year for 30 years (=6000% growth from base) and doesn't invent some unforeseeably radical new technologies, it seems unlikely to me it could cause extinction with <0.1% of its resources.
>
> if AI is doing something equivalent to boosting GWP by 30% per year for >30 consecutive years (>200,000% growth from base), and/or is able to harness arbitrary amounts of nuclear power or solar power / precisely alter molecular structures, it seems fairly likely it could cause extinction with <0.1% of its resources.
>
> ^simplifying here by saying 'power controled [sic] by AI' - I can think of a handful of different ways this would look, with varying levels of human interaction.”

Various participants on the difference between 60% and 99.99% of humans dead

(Online forum discussion)

Pascal:
“Causing 60% vs 99.99% of deaths
It seems like there's a common theme: the concerned crowd thinks killing 60% of humanity is similarly likely to killing 99.9% of humanity, and skeptics think it's ~10x more likely.

I'd be interested if this is largely caused by the *intent* or by the *capability* of AI systems.

e.g. does most of the 10x increase in the skeptic crowd come from it being much easier to kill 60% of people rather than 99.99%? Does the similar probability in the concerned crowd come from assuming it will have the capability to do either of these things, and the intent to do either is ~equally likely?”

> **>Blake (28/04/2023 19:23:00)**

“Re: *I'd be interested if this is largely caused by the *intent* or by the *capability* of AI systems.*

Thanks for the question, Pascal.

Both. Yes, I think more scenarios exist in which it would be easier for AI to kill 60% of the population than 99.9%. And I think it less likely that an AI system would kill everyone than kill 60%, largely for the same reasons Ike outlines below.

But it's not just about the intent and capability of the AI systems. It's also about the intent and capability of humans. In that context, in terms of intent, I think it more likely some humans will weaponize AI to kill many rather than all in a war/terrorist attack. In terms of capability, I think it more likely that 40% of humanity will be capable of surviving to 2100 than none or close to it.

Finally, it's also about timing. There are plenty of scenarios one could cook up that would have an extinction event beginning prior to 2100 but with 60% of the population still alive by the turn of the century, even if their eventual extinction was assured.”

> **>>Pascal (02/05/2023 13:39:00)**
> “Thanks, these seem to me like good points. I'm particularly interested in your final point on eventual extinction which starts with 60% this century - I don't think digging into this would generate promising resolvable-in-5-years cruxes, but it might highlight some of the different views. [O]ne forecast that could maybe get at this is if we included something like p(catastrophe by 2500) in a round of forecasts at some point. My guess is the concerned crowd thinks we're in an unusually risky century ('time of perils'), whereas the skeptics think the risk stays relatively constant each of the next few centuries ('relative' when compared to the concerned group).”
>
> > **>>>Blake (08/05/2023 01:00:00)**
> > “Timing definitely is an issue. I know on another platform, pro GJI superforecasters are far higher for extinction by 2200 than for what we are here for 2100 (I'm at 4%)--still not at the level of the risk concerned, but much closer.”

>Ike (28/04/2023 12:42:00)
“Both sides have agency and humans have many millennia of figuring out how to survive. Humans that are highly centralized and technology dependent may be more vulnerable and considered a greater risk to an AGI/ASI. Tracking down and destroying all humans seems like a waste of time to something that is "intelligent" enough to destroy 60% of us. Why would such a thing fear the remaining 40%, especially if it was continuing to rapidly grow in intelligence and power?

I view that all out nuclear war with the current stockpiles is unlikely to kill 99.99% of human beings but highly likely to kill 60%. The same seems likely with an engineered virus or chemical weapon. The planet is pretty large and communication is good

enough that there is a high probability that we wouldn't all be killed instantaneously. This would give some people the chance to isolate. I don't take it as a foregone conclusion that an AGI or ASI will have the capacity or desire to kill all humans."

>>Pascal (28/04/2023 17:53:00)
"Nice - thanks for this! So it sounds like you think AI is significantly more likely to have the capability AND the intent to kill >60% than >99.99%.
And roughly sounds like you think both capability and intent are of similar importance for your forecast?"

>>>Ike (28/04/2023 18:07:00)
">60% should always be more likely than >99.99% since >99.99% is included in >60%. The capability is essential to reach the number, the intent is not, but is important. An example would be if an[] AGI/ASI didn't particularly care whether humans lived or died and was creating some chemical or biological tool, if it happened to kill us. This may be akin to the extinction of many species that humans have caused without intent."

>>>>Pascal (28/04/2023 19:28:00)
"This is not wrong - maybe I should have operationalized 'significantly' as '>2x'.
I could imagine the likelihood of AI having the intent to kill >60% of people being >2x higher than >99.99%, but I'm still struggling to see that level of likelihood increase for capability (because I expect anything with capability to kill >60% of people will be smart enough to develop more advanced technology and wait until it is certain in capability of killing everyone - assuming here that intent is fixed constant)."

>>>>>Ike (28/04/2023 20:35:00)
"Why would anything with the capability to kill >60% or > 99.99% see us as a significant threat that needs [to be] eliminated? We have not intentionally done that with any species. No species is a threat to our existence. The AI would be getting exponentially more intelligent and powerful. Why spend the energy to destroy every last human? I think that is giving too much credit to our power and importance to such an advanced entity. Why are you confident it would be seeking destruction? What are all of the steps needed to guarantee our destruction and what is the chance that it is thwarted or delayed by any of the steps, either through lack of ingenuity and resources, or because there would be active counter measures deployed against it by humans with access to near-peer alternate AI's or because the evolution of the AI was not rapid enough to avoid being destroyed by humans based on

our need for self-preservation? Each of these[] scenarios should have some probability by which the overall risk is reduced."

>>>>>>Kim (30/04/2023 07:12:00)
"+1. There's a tension where the more powerful the AGI is, the less it needs to exterminate us. The less powerful it is, the more it needs to exterminate us. This leads me to believe that extermination efforts will have a decent chance of failure, as the systems more likely to succeed will be less likely to do it."

>>>>>>Pascal (28/04/2023 22:11:00)
"Oh interesting - I'd been trying to hold intent constant, but it sounds like you think *intent* to kill >60% of people would decrease as *capability* to kill >60% increases.
I think clarifying intent or capability each time is valuable - to avoid mixing the two unintentionally."

>Ash (28/04/2023 12:17:00)
"Capability is the primary reason for the difference in my forecasts on these questions. I view the conditions that must be satisfied to reach 60% deaths as fewer and more probable than those for 99.99%.
But I agree re the intent points brought up by Gus - if there were an AGI/ASI, I can imagine numerous ways that fewer humans would be preferable to no humans. And as well, I do not agree with the assumption that only one side has agency."

>Gus (28/04/2023 11:17:00)
"I made my forecast mostly on capability.
However, I also think intent could come into it. There are scenarios such as the AGSI thinking humans are better off and more self-sufficient with a lower population kept in balance, or the AGSI needs resources we would otherwise consume if our population was larger, or the AGI is testing its capabilities (it would be a big brutal test maybe that gets out of hand) or there is a war that stops when humans give up with so many losses and there is a peace of sorts. There are probably more. Once the thought experiments start, they can go on a long time."

>>Pascal (28/04/2023 17:46:00)
"There are scenarios such as the AGSI thinking humans are better off and more self-sufficient with a lower population kept in balance
This is a good point I hadn't really considered before"

>Dean (28/04/2023 11:19:00)
"10x increase: not me: 2.1% vs 1.5%.

It is partly capability: it is much easier to kill 60% rather than everyone.
It is partly intent: humans can intend to kill 60% other humans, but likely not 100% all humans."

> **>>Pascal (28/04/2023 17:43:00)**
> "Nice, thanks! So you think it's ~40% more likely to kill >60% than to kill >99.99%? That's fairly similar to my prediction (I think mine was ~20%)"

> > **>>>Dean (28/04/2023 17:57:00)**
> > "I am sure our absolute numbers are very different."

> > > **>>>>Hank (29/04/2023 20:21:00)**
> > > "I have probabilities at 60% death at between 4-5x of full extinction. Reasoning is civilizational collapse is a lot more likely and doesn't even necessarily require malevolent AI, if some sort of AI Developed/deployed EMP weapon or genetic modification to a major crop (like rice) goes bad I could see major issues arising, with corresponding chaos causing tons of deaths. As others have noted, we already have the technology to kill 60% of humans currently living. But getting to full extinction requires a leap to much more Sci-FI [sic] type weapons that don't currently exist. I could see the argument for even 10x or greater but I'm hedging a little bit based on the uncertainties if SuperAI were to actually develop."

Wesley<>Blake adversarial collaboration on complex worlds and priors

(Adversarial collaboration call)

Wesley: Imagine that there's a car driving down a highway and the engine keeps getting more powerful. And there's someone driving it, and they can't use the brakes, the accelerator's stuck. They're going faster, and faster, and faster, but they're still steering and in control. And there are other cars on the highway and all they can do is just try to keep going. I feel like there's a disagreement going on where I'm like, wow things are moving faster and faster and faster, and you're still trying to be in control, and you haven't crashed yet, but it's increasingly hard to maintain control because you have to react to stuff as it's happening and your reactions are capped. I can't tell you exactly how that person dies, but I'm pretty sure they end up dead unless they fix the engine problem. And then I imagine some other person (which, in the car case they clearly seem stupid, and in the AI case I don't actually think they seem stupid, but I think this pointing at the disagreement), and they're like: "Look, I know how to forecast stuff: I look at what's happened, and I predict basically the same thing will happen in the future. This person's been driving along the highway for ages and they haven't crashed and they're still steering and they're still looking where they're going and they're not gonna crash because that's the most sensible thing to expect." And I kind of feel like there's a disagreement there, where we can talk about the

tires, or the steering wheel, or the density of cars on the highway. But the actual disagreement is like: is this a stable situation or unstable one?

Blake: I think that's good, because I'm less convinced, especially when you're talking about extinction, that all the microdetails we're talking about measuring are particularly relevant to extinction. I do think that the car analogy is valid to a point. But I think when you're predicting extinction, you're predicting not just that the car is going to crash—which I wouldn't be surprised if it did—you're predicting that it's gonna crash between mile marker 1 and 1.1. You're predicting a very slim sliver of time that it's gonna crash, and with high probabilities. And I think that Karnofsky addresses that a little bit, why this could be a very important century and stuff. But to my mind, it almost has a pre-Copernican idea, where our lives and our time are at the center of the universe. I think the fact that the question resolves in 2100, which is, not coincidentally, just about the span of a human lifetime, people who are alive now will be alive then—these are all kind of baked into the bias of the the question. So, while I do think it's a good thing to look at, I don't think it would be complete. When you ask an extinction question like this, in geologic time you're asking someone to hit a bullseye in a tiny dot.

Moderator: Can I ask a clarifying question? Is it that you think that it's unlikely that we get AI that's powerful enough to cause extinction within such a short time window? Or you think that we may well get AI that's that powerful, but that you expect it to take much longer than 100 years for AI to cause human extinction because you think causing extinction is really hard.

Blake: I don't know. Either/or. I don't rule either out. I think It's entirely possible that we could get AI that could theoretically cause a complete extinction. I mean, it would be hard. People dream up the whole virus—you know, it's disturbing when you see AI come up with different formulas for nerve gas. So it's not out of the question that AI could come out with some super virus or nanobot or something. But even if a super virus were 99.9% lethal and just a few people hung on for another 20 years or so, that wouldn't even come close to resolving the question.

Wesley: Isn't one of the questions cap on GDP growth as well? So I guess a bunch of hunter-gatherer tribes surviving in Papua New Guinea plausibly resolves this existential catastrophe question even if there's more than a million of them around because their GDP is going to be lower.

Blake: Yeah. It's for a million years, though, and once you get to the "less than a million for a million years" thing—they'd have to remain disempowered for a million years which I'm like, maybe, you know, okay.

Wesley: I wonder—there's already questions about "worst catastrophe ever, but very far short of extinction" on there.

Blake: Yeah, and I'm far higher on that. But not anywhere near—

Wesley: What's far higher? Like 10%, 50%?

Blake: 1.5%, up from .1%

Wesley: 1.5%. So still most of the disagreement isn't coming from how hard are humans to kill, most of the disagreement is coming from how soon do we get something that could?

Blake: It's a mix. I'd have to consider where one ended and the other began. But I think people overestimate [underestimate] how difficult this would be, given how diverse humans are politically, in terms of nation states, how diverse we are in terms of body makeups, in terms of how we react to disease and how we react to threat. And that all these things add a layer of complexity, and complexity also adds time. And you're talking about an incredibly short window. It strikes me as indicative of bias that we would be focused on this tiny window that happens to correspond to our lives. Not to say that I'm blind all this stuff that's happening. Another thing: you were getting deep into the specifics of forecasting, x y, or z in terms of the pace of AI development. I do think that's important in that it gives us a way to assess changes in risk, but the ultimate question is more: are we going to commit suicide as a society or not? So it's not about the pace of development of technology so much as whether as a holistic world society, we're going to allow this to happen. I guess how fast this is developing is a proxy of sorts for that, but I think it's a very limited proxy. We're focusing a little bit on regulations and all that stuff, but just how this unfolds—I think there's a danger of focusing too much on just the technological advances because ultimately this is a decision that is being made now by humans, and will be made by humans. And that will involve a lot of political structures and regulation and all that. I didn't phrase that well, but I think it's an important point.

Moderator: I want to give you a chance to respond, Wesley, and then I want to move towards summarizing where we got to, because we're almost out of time.

Wesley: I'm finding this a little difficult to parse in some ways, because I'm not sitting here going "I think there's a 50/50 chance humanity survives the century" and *then* I've decided that AI is the way. I think it requires a really extraordinary thing for humanity to get knocked off track. The sort of extraordinary thing that looks like a species which can do the same kind of things, in our view, as we can for like chimps or dolphins or ants. And then if I see evidence of that kind of extraordinary thing happening, I'm like, okay, I want to talk about the specifics rather than about how big a deal it is for humanity to have been around for this long. I'm not particularly wedded to the century thing. If I try to compress everything into like an incredibly short space, I just want to say something like: if an alien species just appeared on Earth that cognitively was just miles ahead of us and wanted totally different things—maybe breathes carbon dioxide rather than oxygen. So, just doesn't really care about us, but by default just wants a different atmosphere, oxygen is poisonous to them or something. If this happens, are you like "Yeah, humans are all gonna die", or are you still in this position where you're like "Well, you know, I still want to mostly base my forecast —" I'm like, what's bad enough? Because if you're like "even in that world, I just think humans are hard to kill and they'll exist for ages," then I think the disagreement is that. Whereas if you're like "Oh no, that's that's the thing; sure, I'd be terrified in that world, but this AI thing is different," then I think we have to talk about why we have different views on whether AI appears.

Blake: I agree about that. What are you asking there?

Wesley: I still am in this position where I asked, and actually [Moderator] asked, is the disagreement coming from: "We might get AI that is way more cognitively capable than humans, but it'll be fine," or "Why would that happen this century? But sure, if it happens later it'll be really bad." These positions just don't seem that consistent with each other.

Blake: I don't discount either scenario.

Wesley: I'm struggling to reconcile—both of these things seem like positions you hold to some extent, and your one-in-a-thousand chance is—these three things don't seem compatible to me. I really don't understand how to hold all three of them in my head.

Moderator: A part of me really wants to continue on this confusion because it feels productive and will clarify stuff, but we are almost out of time. I think what I want to do with the last five minutes, if it's alright with you guys, is ask each of you to say what you currently expect to be the most promising areas for cruxes between you and the other person? I think it's possible that you'll say different things and that'll be quite interesting, because it feels like we're still figuring out each other's views and stuff.

Wesley: Are we specifically trying to find cruxes that we could turn into forecastable questions? Because I actually think most of the disagreements here are not resolvable by forecasts.

Moderator: I think, don't be too limited to operationalizable cruxes. I'm more just curious for your sense of where the heart of the disagreement between you both is.

Blake: I had one that I'll throw out. In that debate, which I think you [Wesley] read between LeCun and Russell and stuff, LeCun was saying that he thought we'll build safeguards into existing systems and that would be adequate, whereas Russell, in terms of addressing instrumental convergence, Russell was saying that we need to build different systems. That the existing systems were essentially broken and that we need to be thinking about it completely differently. There might be a crux question there: will it be concluded by x date that the current reinforcement learning paradigm that we're going through now that could lead to instrumental convergence problems has been supplanted or agreed to by x number of people. Would that change your [mind]?

Wesley: I think there's a version of that that could work. As it happens. I think Russell and LeCun are both wrong, and this is actually something like the consensus position among current ML researchers. The two problems that are the central part of the essay are 1. Do we solve scalable oversight? and 2. Do we solve deceptive alignment? 95% of my doom is like, if we solve these things, I think we're fine. If we don't solve things, I think we're like 50/50 to be fine, because maybe we'll just get lucky and things will work out that we didn't need to solve problems. There's certainly some questions around how likely do we think it is that people will invent the safeguards that are required in time. Or maybe the disagreement is like, are current safeguards sufficient to steer powerful systems. Does one of those feel more like the crux than the other? Do you think currently our safeguards basically work, or do you think currently, they basically don't but like, maybe we'll invent them in time?

Blake: I don't know enough about the specifics to weigh in definitively, but they don't seem at a glance to me to be sufficient, and I don't know that we'll invent them in time. Which may seem inconsistent with my forecast, but that, my, well—

Wesley: I think I am just still pretty confused about this because I'm like: I agree that we might not need them, I agree that it doesn't seem like we will invent them in time, and I just don't get one-in-a-thousand out of that.

Blake: Well, I mean, I say I just don't *know* if we'll invent them in time. I suspect it will be haphazard and partially effective, if I had to guess.

Moderator: I think we should move towards wrapping up, which is unfortunate because this is interesting. I'm kind of unsure what the confusion is here. I wonder if there's some pretty deep worldview difference thing where Blake is just paying a lot of attention to complexity and is like "well, all these things are possible, so are a million things I haven't even thought of, which eat up a lot of my probability mass." And then Wesley, you're looking at the specific things that seem like a large chunk of your probability mass, and it seems inconsistent under that model, but it wouldn't be in this different worldview where there are millions and millions of unknown things or something.

Blake: Maybe that's a fair characterization.

Wesley: I think that's close to it. I think that the disagreement isn't that I am only looking at specific things. I'm like: Why does the complexity save you? I feel like what's going on is: "oh, things could turn out in a way that we didn't predict." And I'm like, "why are all of the ways we didn't predict good?" Not all of the ways. But why are 99% of the way we didn't predict, why did humans end up fine in all of those? If you go to the maximally—if you just rearrange all of the atoms in the universe randomly, humans aren't alive in most of those rearrangements. If you run evolution a bunch of times, probably you get some stuff that looks pretty different to humans. There's some uniform prior where if you're just like "oh, the world's super complex, I can't predict anything," I think mostly you just get a bunch of random gas. So we have to do some prediction based on stuff we're looking at. So why is the point where we stop, like, "yeah, all of these risks do seem plausible, but then I'll just put a bunch of mass on 'in some way I can't actually tell a story for, humans end up fine.'" I realize that's a straw man. But that's what it feels like, and that's why I keep getting confused.

Moderator: It feels like it comes back down to the general prior thing of, by default, do we expect everything to be fine or by default with everything to be bad? I feel like in this call we haven't got deep enough in that to quickly resolve anything. I think we should wrap up because I don't want to eat into your time. I really enjoyed that and thought it was actually a really productive exchange of quite different views. I still feel confused.

Blake: I still think we need to get cruxes out of it. I've enjoyed it too and it gives me a better sense of where Wesley is coming from and I'll try to think more about it. It's a little like pinning jello to the wall for me, finding something that would really move the needle, but I'll give it more thought.

Wesley: I think it might be good for us to try to—[Moderator], I think, you had a transcript last time? I think seeing if we can get some kind of comment thing going basically about—I think the most important thing is this worldview, what's the prior: Does complexity make things easier or harder? I think trying to dig into that does seem like the most useful place to converge, because I just feel like at the object level—yeah, this seems to be accounting for most of the difference in prediction.

Moderator: Yeah, after the call I'll share with you guys the notes I've made and the transcript in case that's helpful for you thinking further about this. And I'll also share my version of a summary of this call for the forum, because I feel a bit uncertain, I would like to check that my summary is not totally off. But yeah, thanks so much both for your time today. I think that was pretty good progress.

Relevant Wesley quotes from elsewhere in the project:

- “I think there is maybe some meta disagreement, where you say, “there are loads of ways this could go—why are you so worried about the bad ways?” And I say, “there are loads of ways this could go and very few of them leave humans alive.”
- "I think sticking close to reference classes is less appropriate in this domain—I'm making object level arguments instead of reference classes because I think the reference classes are doing less work than they typically do for forecasts like this.”

Appendix 9: Directions of updates

Question	Group	Number who would be more concerned if question resolved "yes"	Number who would be less concerned if question resolved "yes"	Number who would not update based on whether question resolved "yes"	Total number of participants who forecasted on this question
6 month pause	Concerned	0	7	1	8
	Skeptical	2	2	5	9
AI Forecasting skill	Concerned	4	0	4	8
	Skeptical	3	1	5	9
AI Robotics	Concerned	4	0	1	5
	Skeptical	2	0	9	11
AI articles and apps	Concerned	7	0	1	8
	Skeptical	3	1	5	9
AI coding	Concerned	5	0	0	5
	Skeptical	5	0	6	11
AI solving novel math problems	Concerned	7	1	0	8
	Skeptical	3	0	6	9
AI writes AI	Concerned	8	0	0	8
	Skeptical	9	0	1	10
Alignment researchers changing minds	Concerned	0	8	0	8
	Skeptical	1	2	6	9
Alignment solution	Concerned	0	8	0	8
	Skeptical	0	7	3	10
Cyberattacks	Concerned	2	3	3	8

	Skeptical	4	0	6	10
Democratic influence	Concerned	5	3	0	8
	Skeptical	9	0	1	10
Escalating warning shots	Concerned	6	2	0	8
	Skeptical	6	1	1	8
Evidence of misalignment	Concerned	5	2	1	8
	Skeptical	5	0	2	7
Fast AI efficiency gains	Concerned	7	0	1	8
	Skeptical	5	0	4	9
IC demonstration	Concerned	0	3	5	8
	Skeptical	4	0	6	10
IT progress	Concerned	0	5	0	5
	Skeptical	2	8	1	11
Intergovernmental AI safety	Concerned	6	0	2	8
	Skeptical	3	0	7	10
Muehlhauser policies	Concerned	4	1	0	5
	Skeptical	4	4	3	11
Major powers war	Concerned	0	7	0	7
	Skeptical	0	6	4	10
No violence LLM	Concerned	0	8	0	8
	Skeptical	0	2	7	9
Non-democracy AI	Concerned	6	0	2	8
	Skeptical	3	1	5	9

Other fields IC	Concerned	1	1	3	5
	Skeptical	5	0	6	11
Platform: AI regulation	Concerned	0	11	0	11
	Skeptical	0	6	5	11
Platform: ARC Evals	Concerned	8	3	0	11
	Skeptical	9	0	2	11
Platform: Escalating warning shots	Concerned	3	4	0	7
	Skeptical	8	2	1	11
Platform: Transformative growth	Concerned	5	6	0	11
	Skeptical	5	2	4	11
Politicization	Concerned	2	0	3	5
	Skeptical	1	1	9	11
Power-seeking	Concerned	5	1	1	7
	Skeptical	7	1	2	10
Power-seeking shutdown	Concerned	3	1	1	5
	Skeptical	9	0	2	11
Progress in lethal technologies	Concerned	3	1	4	8
	Skeptical	5	0	2	7
Public concern	Concerned	2	3	0	5
	Skeptical	7	1	3	11
Reduction in AI investment	Concerned	1	6	1	8
	Skeptical	1	5	3	9
Req testing	Concerned	0	2	5	7

	Skeptical	0	3	7	10
Short-term GDP change	Concerned	4	0	4	8
	Skeptical	2	0	7	9
Supers changing minds	Concerned	4	0	2	6
	Skeptical	5	0	2	7
Taiwan-China	Concerned	4	2	2	8
	Skeptical	3	1	6	10
Warning shot	Concerned	7	1	0	8
	Skeptical	5	2	1	8

Table 23. For each question, the number of participants who would update their P(AI Extinction by 2100) upwards if the question resolved "yes," the number who would update downwards, and the number who would not update either way (i.e. they believe the question bears no relevance to extinction risk).

Appendix 10: Areas of disagreement

Most of the written disagreement on P(AI X-risk by 2100) clustered in these four areas:

1. AI will/not develop sufficiently in the time frame
2. Goals that incentivise killing everyone are very rare/common
3. Killing everyone will/not remain really hard
4. Responses will be in/adequate

See Timelines for AI Progress and Goals that incentivize killing everyone for detailed discussion of 1 and 2. See below for discussion of 3 and 4. And see "Understanding each other's arguments" for an alternative description of key arguments and areas of disagreement.

Killing everyone will/not remain really hard

Main arguments from the skeptic group:

- It seems very unlikely that current technologies like nuclear weapons or biotechnology would be sufficient for killing all humans.[191]
- Technologies like Dyson spheres and advanced nanotechnology are unlikely to be developed by AI in the relevant timeframe.[192]
- Even with future weapon technologies which could theoretically wipe out all humans, the logistics required would be enormous.[193]
- Humans are resilient and distributed.[194]

Main arguments from the concerned group:

---

[191] "I view that all out nuclear war with the current stockpiles is unlikely to kill 99.99% of human beings but highly likely to kill 60%. The same seems likely with an engineered virus or chemical weapon."(Gus).; "…and there has to be a credible method of causing extinction - not a catastrophe but extinction - most people on the [XPT] project felt nuclear war, pathogens, asteroids etc...were unlikely to to [sic] do so." (Gus).

[192] "An actual Dyson sphere in the next 77 years? Yes, put me down in the extremely skeptical column!" (Blake).

[193] "Obviously, I think it's possible to come up with new ideas to cause extinction. I think the ability to execute those ideas is equally important, and I wonder (a) if this is the more complex and uncertain part of the question and (b) whether the computational power is as relevant to that part of it (which might be where we disagree). I think it will still be very hard to kill everyone quickly in a way that humans could not respond to)." (Gus).

[194] "Remember, humans are distributed across the planet, in some quite remote and inaccessible places, with diverse genetics… " (Gus); "The planet is pretty large and communication is good enough that there is a high probability that we wouldn't all be killed instantaneously. This would give some people the chance to isolate." (Ike); "In the capital city of one of Canada's territories (Iqaluit, Nunavut), the population is approaching 10k and you can only access it by air or sea. If an AI event takes down most of the world, is Iqaluit going to be susceptible? Almost certainly not." (Claire); "~75% of the population of PNG live in rural areas on a subsistence level. That means that if just those people survived, there would be more than 7M people remaining on earth." (Ike); "Since humans are adaptable, widely distributed, unlikely to just sit in place and die, this change would have to take place extremely fast." (Ash); "Were something to go terribly wrong—and it very well might by 2100—we're not exactly defenseless" (Blake); "There are 8 billion people on earth with a head start, vast militaries and intelligence communities as well as corporations, that all want to survive and prosper. If they view AI as a threat, there is a vast potential for cyber as well as kinetic methods to thwart misaligned AI." (Ike).

- An ASI would have the capacity to radically alter the world to overcome the barriers to human extinction (e.g. by making it so that humans were less distributed and by developing new means of killing on a large scale).[195]
- Our inability to imagine methods for causing human extinction is not much grounds for optimism. An ASI would be much smarter than us and we should expect it to devise scenarios that we can't.[196]

Responses will be in/adequate

Main arguments from the skeptic group:

- By default, a lot of attention will be paid to the safety of AI systems and we should expect guardrails to be put in place.[197]
- Governments are usually effective at regulation.[198]

---

[195] "My base expectation is that, if an existential catastrophe happens, it happens by an AI building up from 15% annual growth to, one-ten years later, >100% annual growth to, one-ten years later, capacity to cause existential catastrophe. e.g. by building up enough public favor to support development of a world such that there are very few people living inaccessibly, and those people are still significantly less inaccessible than people are now. And having excellent details about their whereabouts nonetheless." (Pascal); "I expect causing extinction will be relatively very easy and low-cost for a very advanced technological civilization that doesn't care about consequences of extinction. The kind of thing that the equivalent of a small lab could do if it wanted to." (Riley); "Destroying the world is offense-dominant: I think it may not be terribly difficult for an AI to wipe out most or all of humanity - i.e. I think an AI which is still subhuman in many important domains (e.g. locomotion) could still wipe out humanity. E.g. If an AI is superhuman at finding and exploiting zero day exploits, has some degree of agency (or mandate), and can self-replicate - this already seems incredibly destructive. If it has any ability to use the compute it acquires in this way to further fine-tune itself or other-wise self-improve then you have a system which can essentially dominate the world. Add in the ability to create bioweapons (e.g. using cloud labs or by manipulating humans into sending dangerous pathogens) and you have a system which can easily wipe out humanity. It need not be particularly smart except in the domain of cyber (but it need not be far beyond humans here either) and it already has a path to destruction." (Vincent).

[196] "And maybe it will do all of these and a hundred more simultaneously, because it is a superintelligence that has gotten access to a ton more compute via the internet, and I am just a human and can't come up with that many strategies that would actually kill all humans." (Ume).

[197] "As with any software system, anything critical (read: "not a chatbot that can be asked for a hypothetical scenario of eradicating Sudan") will be severely tested by testers, ethical hackers, and malicious hackers including the ones working for nation states" (Anonymous Skeptic); "Since ChatGPT and GPT-4 came out, I have updated higher on the chances of governments regulating this technology. They are paying attention, and they do care. The more powerful AI becomes, the more everyone in the world will care." (Kim) "It feels plausible to me that we train something with many safeguards and monitoring, like not using deception/violence/power seeking/etc." (Kim); "Further, strong AI with major systems access will probably not exist for a significant amount of this time. Also, it seems unlikely to me that we will not have any mitigants or controls to make this less likely." (Flint); "Humans are very adaptable, and will put up sufficient guardrails. That will include AIs that search for mis-aligned AIs (and AIs which search for mis-aligned AIs that are supposed to search for mis-aligned AIs ...). I think elsewhere I have used the analogy of an immune system, where human beings are the killer T-cells, antibodies, etc." (James).

[198] "Self-driving cars as an example of where, I think, regulatory bodies have slowed down progress in an interesting way. So I do expect that if there are warning shots, which may happen if the AI starts to be very useful, that there is going to be a real clamp down in how governments, which are very risk averse, address these issues." (FRI moderator, call with Zoe).

- AI advancements will take place in dynamic systems, and when AI begins to have economic impacts or harmful impacts, our social and political systems will adjust to limit negative effects.[199]
- There is a warning paradox, where the more people are worried about AI risk, the more likely it is that serious responses are made.[200]

Main arguments from the AI risk concerned group:

- Previous efforts at global regulation (e.g. to reduce nuclear risk, limit climate change) have not been sufficiently effective.[201]
- Regulation usually occurs in response to an incident, rather than preceding one.[202]
- There are incentives that push away from caution, and there will likely be a greater number of less cautious actors in the future.[203]

[199] "A number of proposed cruxes pertain to an AI doing something bad by 2030, suggesting there's a higher chance of an AI doing something a lot worse in future years. OTOH something bad is going to trigger humans to do something about it. Essentially humans act like an immune system. So for me to increase my forecast, it's not sufficient that an AI does something bad." (James).

[200] Another somewhat related issue might be the warning paradox. My relatively low forecast is in part due to the relatively high forecasts of those who do AI safety research. The reverse could also be true; their relatively high forecast may in part be due to my relatively low forecast. So IMHO it's interesting (and perhaps unwise) that I'm more confident than the people actually doing the research that they will make progress. (James).

[201] "We've come very close to reaching nuclear war on multiple occasions. E.g. Cuban missile crisis. Do people think the probability of nuclear war during the cuban [sic] missile crisis was 0.1%? Nuclear launch passwords were set to 0000 for a long time, just to illustrate. Experts seem to think nuclear safety is very inadequate. Nuclear proliferation has happened despite attempts to stop it, though it has had some effect. The government response to climate change has been pretty bad, despite widespread belief (possibly wrongly) that this will cause extinction." (Riley).

[202] "Regulation usually happens years after the first problems arise." (Riley).

[203] "[T]he incentive landscape will push heavily toward building more agentic systems and giving them a broader ability to act without human supervision. Also, I worry that models will proliferate quickly to less-responsible actors." (Vincent); "Economic and political incentives to building AI systems AND giving them material resources are sufficiently strong that, even if any given individual company has some small x% subjective probability of being the one that generates a catastrophe, they will be inclined to take the chance if they have a much larger than x% probability of gaining large wealth from it (individualized rewards, systemic risk), assuming there is no coordinated attempt to solve the coordination failure." (Riley).