
THOUSANDS OF AI AUTHORS ON THE FUTURE OF AI

PREPRINT

Katja Grace^{*†}
AI Impacts
Berkeley, California
United States
katja@aiimpacts.org

Harlan Stewart[†]
AI Impacts
Berkeley, California
United States

Julia Fabienne Sandkühler[†]
Department of Psychology
University of Bonn
Germany

Stephen Thomas[†]
AI Impacts
Berkeley, California
United States

Ben Weinstein-Raun
Independent
Berkeley, California
United States

Jan Brauner
Department of Computer Science
University of Oxford
United Kingdom

January 2024

ABSTRACT

In the largest survey of its kind, we surveyed 2,778 researchers who had published in top-tier artificial intelligence (AI) venues, asking for their predictions on the pace of AI progress and the nature and impacts of advanced AI systems. The aggregate forecasts give at least a 50% chance of AI systems achieving several milestones by 2028, including autonomously constructing a payment processing site from scratch, creating a song indistinguishable from a new song by a popular musician, and autonomously downloading and fine-tuning a large language model. If science continues undisrupted, the chance of unaided machines outperforming humans in every possible task was estimated at 10% by 2027, and 50% by 2047. The latter estimate is 13 years earlier than that reached in a similar survey we conducted only one year earlier [Grace et al., 2022]. However, the chance of all human occupations becoming fully automatable was forecast to reach 10% by 2037, and 50% as late as 2116 (compared to 2164 in the 2022 survey).

Most respondents expressed substantial uncertainty about the long-term value of AI progress: While 68.3% thought good outcomes from superhuman AI are more likely than bad, of these net optimists 48% gave at least a 5% chance of extremely bad outcomes such as human extinction, and 59% of net pessimists gave 5% or more to extremely *good* outcomes. Between 37.8% and 51.4% of respondents gave at least a 10% chance to advanced AI leading to outcomes as bad as human extinction. More than half suggested that “substantial” or “extreme” concern is warranted about six different AI-related scenarios, including spread of false information, authoritarian population control, and worsened inequality. There was disagreement about whether faster or slower AI progress would be better for the future of humanity. However, there was broad agreement that research aimed at minimizing potential risks from AI systems ought to be prioritized more.

*Corresponding author

†Equal Contribution

1 Introduction

Artificial intelligence appears poised to reshape society. Decision-makers are working to address opportunities and threats due to AI in the private sector [OpenAI, 2023], academia [Center for Human-compatible Artificial Intelligence, 2023], and government at the state, national, and international levels [Newsom, 2023, AI.gov, 2023, Inter-Agency Working Group on Artificial Intelligence, 2022].

Navigating this situation requires judgments about how the progress and impact of AI are likely to unfold. However, there is a lack of apparent consensus among AI experts on the future of AI [Korzekwa and Stewart, 2023]. These judgments are difficult, and there are no established methods of making them well. Thus we must combine various noisy methods, such as extrapolating progress trends [Villalobos, 2023]; reasoning about reference classes of similar events [Grace et al., 2021]; analyzing the nature of agents [Omohundro, 2008]; probing qualities of current AI systems and techniques [Park et al., 2023]; applying economic models to AI scenarios [Jones, 2023, Trammell and Korinek, 2023]; and relying on forecasting aggregation systems such as markets, professional forecasters, and the judgments of various subject matter experts.

One important source of evidence comes from the predictions of AI researchers. Their familiarity with the technology and the dynamics of its past progress puts them in a good position to make educated guesses about the future of AI. However, they are experts in AI research, not AI forecasting and might thus lack generic forecasting skills and experience, or expertise in non-technical factors that influence the trajectory of AI. While AI experts' predictions should not be seen as a reliable guide to objective truth, they can provide one important piece of the puzzle.

We conducted a survey of 2,778 AI researchers who had published peer-reviewed research in the prior year in six top AI venues (NeurIPS, ICML, ICLR, AAAI, IJCAI, JMLR). This to our knowledge constitutes the largest survey of AI researchers to date. The survey took place in the fall of 2023, after an eventful year of broad progress (including the launch of ChatGPT and GPT-4, Google's Bard, Bing AI Chat, Anthropic's Claude 1 and 2), shift in public awareness of AI issues (including two widely signed and publicized AI safety letters [Future of Life Institute, 2023, Center for AI Safety, 2023]), and governments beginning to address questions of AI regulation in the US, UK, and EU [Biden, 2023, Amodi, 2023, gov.uk, 2023, European Parliament, 2023]. The survey included questions about the speed and dynamics of AI progress, and the social consequences of more advanced AI.

2 The Survey

This survey, the "2023 Expert Survey on Progress in AI," or ESPAI, is the third in a series of very similar surveys. The first two were conducted in 2016 [Grace et al., 2018a] and 2022 [Grace et al., 2022]. The 2023 survey included around four times as many participants as the 2022 survey by expanding from two publication venues (NeurIPS and ICML) to six. It also includes several new questions to probe the nature of future AI systems and diverse potential risks. The survey complements a collection of recent work gathering views on similar questions from the public [Stein-Perlman, 2023] and corporate leadership [Chui et al., 2023]. The full set of questions is available from AI Impacts [2023a].

Most questions solicited responses in one of three ways: on a Likert scale (multiple choice along a single axis); a probability estimate; or an estimate of a future year. A smaller number of questions asked for write-in responses or numerical estimates.

The way a question is framed can greatly influence the response, and this seems more likely for complex questions [Tversky and Kahneman, 1981]. To assess and mitigate framing effects, we often posed different variations of questions on the same topic to different random subsets of respondents. For example, all questions about how soon a milestone would be reached were framed in two ways: fixed-years and fixed-probabilities. Half of respondents were asked to estimate the probability that a milestone would be reached by a given year ("fixed-years framing"), while the other half were asked to estimate the year by which the milestone would be feasible with a given probability ("fixed-probabilities framing"). To minimize confusion, each participant received one framing throughout the survey.

In several parts of the survey, each participant randomly received questions on only one of several topics, to keep the survey brief. We allocated these questions to differently sized subsets of participants based on factors like the importance of the question and the relative value of a larger sample size. This means that most questions were not assigned to all 2,778 participants (see Section 6: Methods).

Several questions asked participants to estimate how many years until a milestone will be feasible. In these questions, we asked participants to provide three year-probability pairs (either via the fixed-years framing or fixed-probabilities framing described above), which we used to approximate a probability distribution for that participant by fitting a gamma cumulative distribution function to these points.

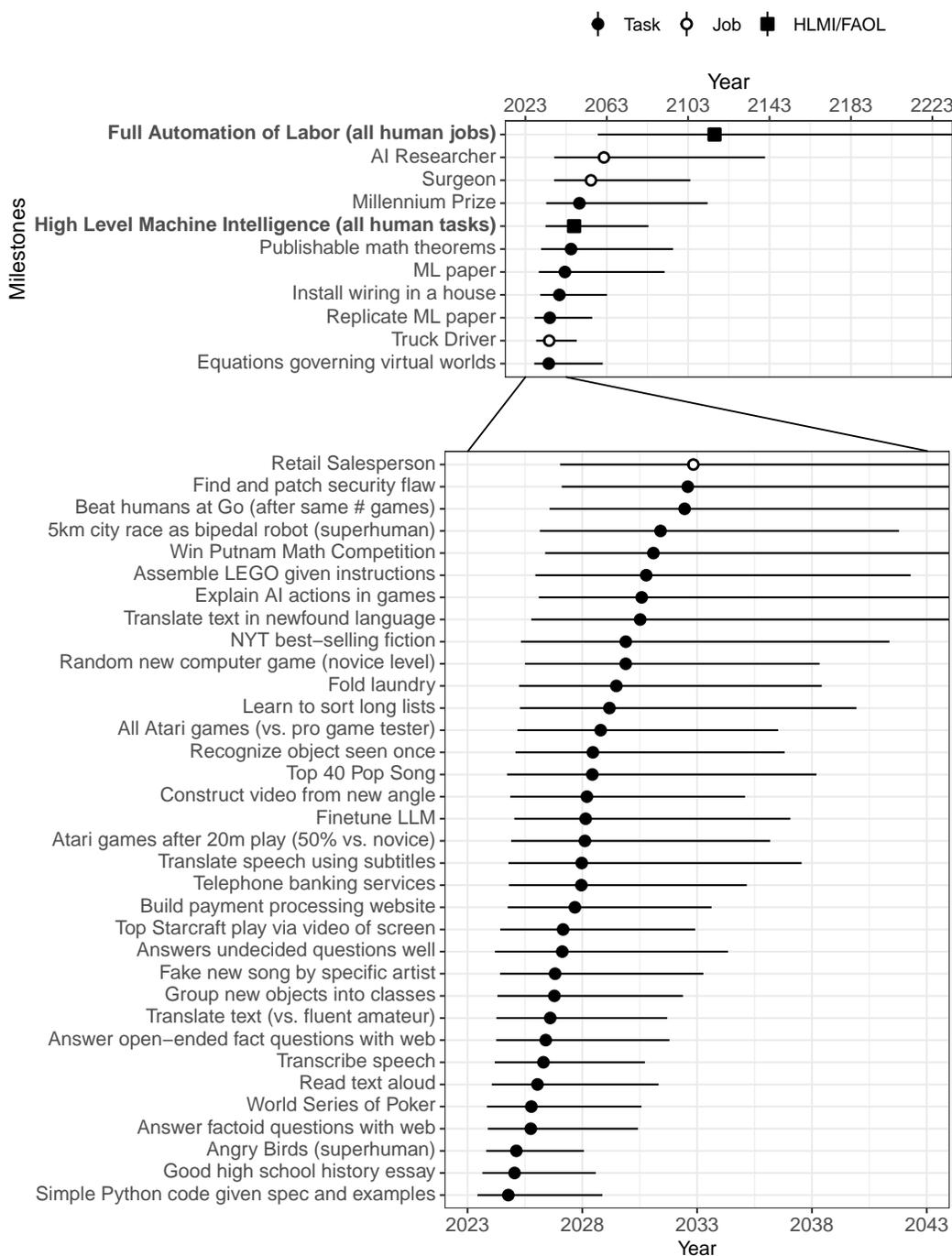


Figure 1: **Most milestones are predicted to have better than even odds of happening within the next ten years, though with a wide range of plausible dates.** The figure shows aggregate distributions over when selected milestones are expected, including 39 tasks, four occupations, and two measures of general human-level performance (see Section 3.2), shown as solid circles, open circles, and solid squares respectively. Circles/squares represent the year where the aggregate distribution gives a milestone a 50% chance of being met, and intervals represent the range of years between 25% and 75% probability. Note that these intervals represent an aggregate of uncertainty expressed by participants, not estimation uncertainty. The displayed milestone descriptions are summaries; for full descriptions, see Appendix C.

3 Results on AI Progress

3.1 How soon will 39 tasks be feasible for AI?

The survey asked about when each of 39 tasks would become feasible, where “feasible” was described as meaning “one of the best resourced labs could implement it in less than a year if they chose to. Ignore the question of whether they would choose to.” Each respondent was asked about four tasks, so that each task received around 250 estimates. Each respondent gave three probability-year pairs per task.

To aggregate the responses, we first fit a gamma distribution to each participant’s three probability-year pairs, and then computed the mean across the participants’ individual gamma distributions.

All but four of the 39 tasks were predicted to have at least a 50% chance of being feasible within the next ten years (Figure 1). This includes several economically very valuable tasks—such as coding an entire payment processing site from scratch and writing new songs indistinguishable from real ones by hit artists such as Taylor Swift. It also includes tasks that imply substantial progress in sample-efficiency (e.g. ‘Beat novices in 50% of Atari games after 20 minutes of play’), AI-driven AI progress (e.g. autonomously fine-tuning an open-source LLM), and robotics (e.g. folding laundry).

The six tasks expected to take longer than ten years were: “After spending time in a virtual world, output the differential equations governing that world in symbolic form” (12 years), “Physically install the electrical wiring in a new home” (17 years), “Research and write” (19 years) or “Replicate” (12 years) “a high-quality ML paper,” “Prove mathematical theorems that are publishable in top mathematics journals today” (22 years), and solving “long-standing unsolved problems in mathematics” such as a Millennium Prize problem (27 years).

3.1.1 Comparison with 2022

32 AI task questions were identical to those in the 2022 survey, as well as the 2016 survey [Grace et al., 2018a], with the exception of minor edits to task descriptions for updated accuracy between 2016 and 2022 [AI Impacts, 2023b]. All tasks from Grace et al. [2018a] were included, regardless of whether the authors would judge them to be achieved. The survey population included more conferences in 2023, but this did not appear to have a notable effect on opinion (See 5.2.3). Figure 2 shows how expected dates for reaching these milestones shifted from 2022 to 2023.

Between 2022 and 2023, aggregate predictions for 21 out of 32 tasks moved earlier. The aggregate predictions for 11 tasks moved later.

On average, for the 32 tasks included in both the 2022 and 2023 surveys, the 50th percentile year they were expected to become feasible shifted 1.0 years earlier (SD = 2.0, SE = 0.18).

3.2 How soon will human-level performance on all tasks or occupations be feasible?

We asked how soon participants expected AI systems to outperform humans across all activities, which we framed in two ways: as either *tasks*, in the question about “High-Level Machine Intelligence” (HLMI), or *occupations*, in the question about “Full Automation of Labor” (FAOL).

3.2.1 How soon will ‘High-Level Machine Intelligence’ be feasible?

We defined High-Level Machine Intelligence (HLMI) thus:

High-level machine intelligence (HLMI) is achieved when unaided machines can accomplish every task better and more cheaply than human workers. Ignore aspects of tasks for which being a human is intrinsically advantageous, e.g. being accepted as a jury member. *Think feasibility, not adoption.*

We asked for predictions, assuming “human scientific activity continues without major negative disruption.” We aggregated the results (n=1,714) by fitting gamma distributions, as with individual task predictions in 3.1.

In both 2022 and 2023, respondents gave a wide range of predictions for how soon HLMI will be feasible (Figure 3). The aggregate 2023 forecast predicted a 50% chance of HLMI by 2047, down thirteen years from 2060 in the 2022 survey. For comparison, in the six years between the 2016 and 2022 surveys, the expected date moved only one year earlier, from 2061 to 2060¹.

The aggregate 2023 forecast predicted a 10% chance of HLMI by 2027, down two years from 2029 in the 2022 survey. See Appendix B for a table comparing this with Full Automation of Labor, which is discussed below.

¹Reported as 2059 in 2022; small code changes and improvements to data cleaning between surveys shifted the aggregate slightly.

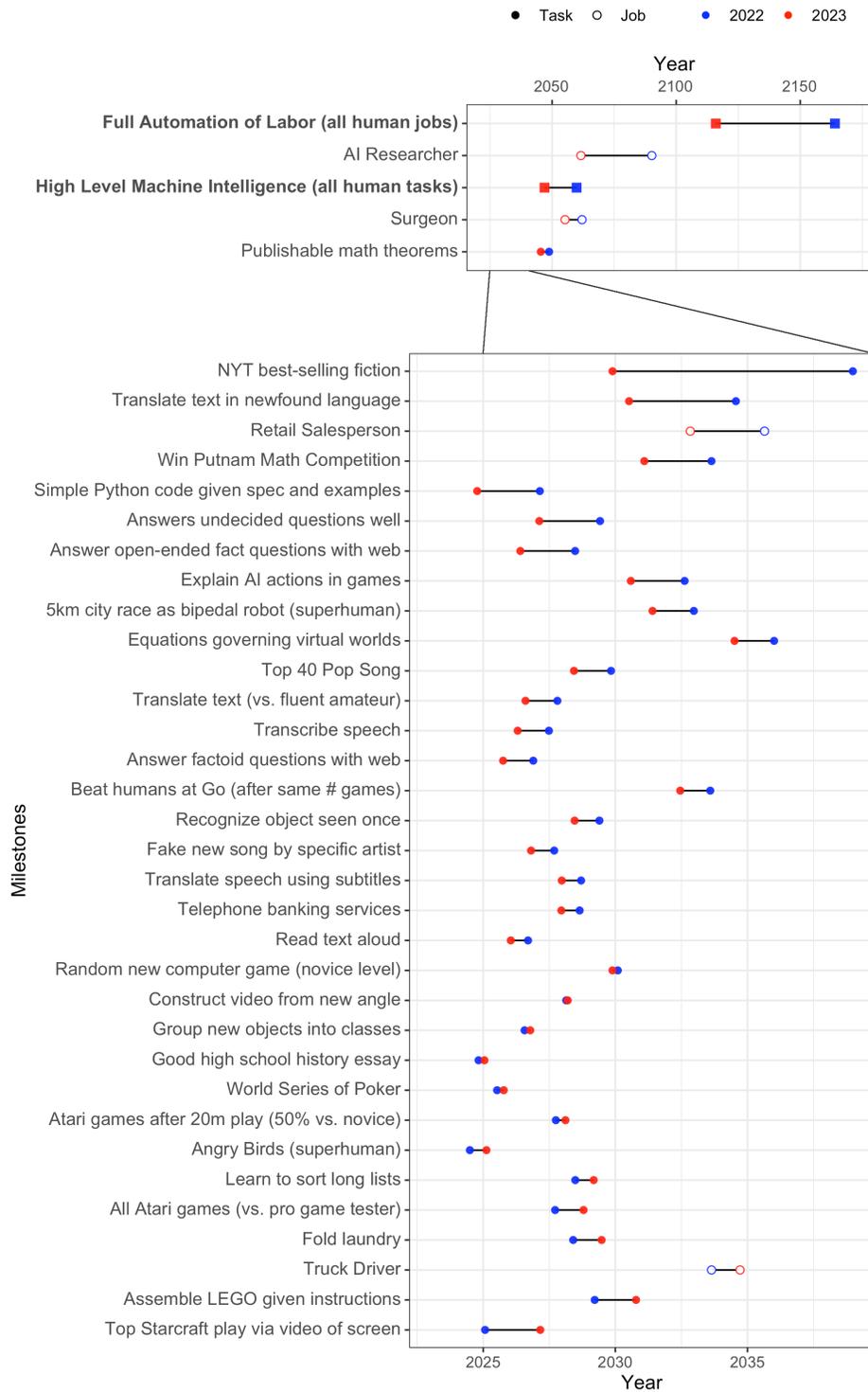


Figure 2: **Expected feasibility of many AI milestones moved substantially earlier in the course of one year (between 2022 and 2023).** The milestones are sorted (within each scale-adjusted chart) by size of drop from 2022 forecast to 2023 forecast, with the largest change first. The year when the aggregate distribution gives a milestone a 50% chance of being met is represented by solid circles, open circles, and solid squares for tasks, occupations, and general human-level performance respectively. The three groups of questions have different formats that may also influence answers. For full descriptions of the summarized milestones, see Appendix C.

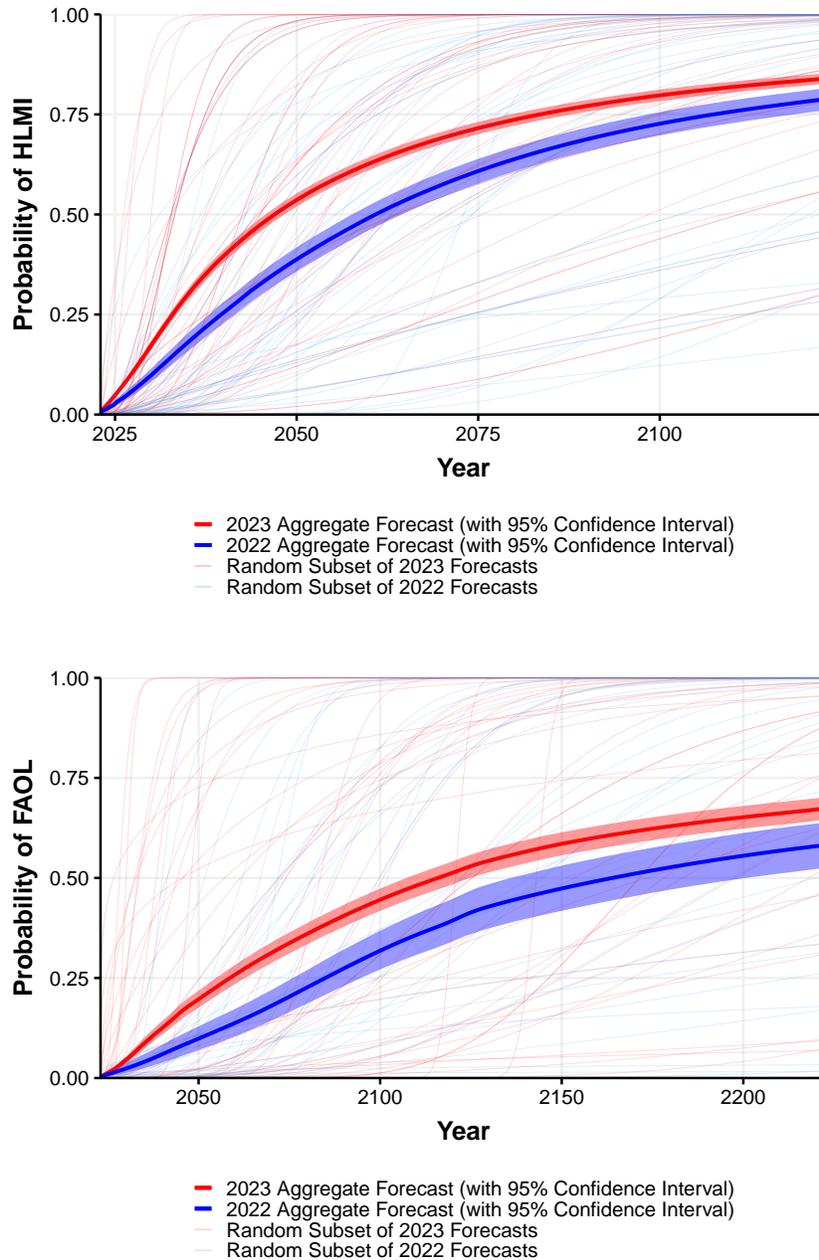


Figure 3: **Aggregate forecast for 50th percentile arrival time of High-Level Machine intelligence (HLMI) dropped by 13 years between 2022 and 2023. The forecast for 50th percentile arrival time of Full Automation of Labor (FAOL) dropped by 48 years in the same period. However, there is still a lot of uncertainty.** “Aggregate Forecast” is the mean distribution over all individual cumulative distribution functions. For comparison, we included the 2022 Aggregate Forecast. To give a sense of the range of responses, we included random subsets of individual 2023 and 2022 forecasts. Note that the thinner ‘confidence interval’ in 2023 (compared to 2022) is due to our increased confidence about the average respondents’ views due to a larger sample size, not respondents’ predictions converging.

3.2.2 How soon will ‘Full Automation of Labor’ be feasible?

The other framing of the question about how soon AI systems would outperform humans across all activities was about “Full Automation of Labor,” or FAOL. We defined FAOL thus:

Say an occupation becomes fully automatable when unaided machines can accomplish it better and more cheaply than human workers. Ignore aspects of occupations for which being a human is intrinsically advantageous, e.g. being accepted as a jury member. *Think feasibility, not adoption.* [...]

Say we have reached ‘full automation of labor’ when all occupations are fully automatable. That is, when for any occupation, machines could be built to carry out the task better and more cheaply than human workers.

Before the participants (n=774) were asked about the full automation of labor, respondents were asked when four specific occupations would become fully automatable: “Truck driver,” “Surgeon,” “Retail salesperson,” and “AI researcher” (Figure 1). They were also asked to think of an existing human occupation that they thought would be among the final ones to be fully automatable. They were then asked when ‘full automation of labor’ (FAOL) would be achieved.

The aggregate 2023 forecast predicted a 50% chance of FAOL by 2116, down 48 years from 2164 in the 2022 survey (Figure 3). We checked if this difference was significant for participants who received the question in the fixed-probabilities framing, and found that it was ($p = .0052$, Yuen’s test (bootstrap version); see Appendix F). There is about a 70-year difference between the mean 50% prediction for HLMI and the mean 50% prediction for FAOL. We discuss this surprising finding in the next section, “Framing effect of HLMI vs FAOL.”

Compared to 2016, 2023 has earlier 50% estimates but later 10% estimates. (See Appendix B)

The answers to the write-in question about an existing occupation likely to be among the last automatable were categorized according to O*NET’s *All Job Family Occupations* categories [O*NET, 2023]. The top five most-suggested categories were: “Computer and Mathematical” (91 write-in answers in this category), “Life, Physical, and Social Science” (77 answers), “Healthcare Practitioners and Technical” (56), “Management” (49), and “Arts, Design, Entertainment, Sports, and Media” (39).

3.2.3 Differences between HLMI and FAOL

Predictions for a 50% chance of the arrival of FAOL are consistently more than sixty years later than those for a 50% chance of the arrival of HLMI. This was seen in the results from the surveys of 2023, 2022, and 2016. This is surprising because HLMI and FAOL are quite similar: FAOL asks about the automation of all occupations; HLMI asks about the feasible automation of all tasks. Since occupations might naturally be understood either as complex tasks, composed of tasks, or closely connected with one of these, achieving HLMI seems to either imply having already achieved FAOL, or suggest being close.

We do not know what accounts for this gap in forecasts. Insofar as HLMI and FAOL refer to the same event, the difference in predictions about the time of their arrival would seem to be a framing effect.

However, the relationship between “tasks” and “occupations” is debatable. And the question sets do differ beyond definitions: only the HLMI questions are preceded by the instruction to “assume that human scientific activity continues without major negative disruption,” and the FAOL block asks a sequence of questions about the automation of specific occupations before asking about full automation of labor. So conceivably this wide difference could be caused by respondents expecting major disruption to scientific progress, or by the act of thinking through specific examples shifting overall anticipations. From our experience with question testing, it also seems possible that the difference is due to other differences in interpretation of the questions, such as thinking of automating occupations but not tasks as including physical manipulation, or interpreting FAOL to require *adoption* of AI in automating occupations, not mere *feasibility* (contrary to the question wording).

3.2.4 Demographic differences

Geographical background was correlated with expectations about the timing of human-level performance: respondents whose undergraduate education was in Asia anticipated an 11 year earlier arrival of HLMI than participants from Europe, North America, or other regions combined. See Appendix A for more demographic comparisons.

3.2.5 Do participants think they agree on timing of HLMI?

We asked respondents a set of “meta” questions about their views on others’ views (n=671). One meta question asked to what extent they thought that they disagreed with the typical AI researcher about when HLMI would exist. 44% said “Not much,” 46% said “A moderate amount,” and 10% said “A lot.”

3.3 Framing effect of fixed-years vs fixed-probabilities

All questions about how soon a milestone would be reached were framed in two ways: fixed-years and fixed-probabilities. In either framing, we ask for three year-probability pairs, but in one we fix a set of probabilities (10%, 50%, 90%) and ask how many years until the participant would assign each probability to the milestone being met, whereas in the other framing we fix a set of future years (usually 10 years, 20 years, 50 years) and ask about the probability of the milestone occurring by that year.

The fixed-years framing has been previously observed to produce systematically later predictions [Grace et al., 2018a, 2022], but we do not know if one framing is more accurate than the other. Here we have used both and combined them with equal weight.

The previously-observed framing effect was again observed in this survey. For example, the year with a 50% chance of HLMI from participants answering in the fixed-year frame (34 years) was twice as far into the future as that for participants answering in the fixed-probability frame (17 years). However, it’s notable that even the larger of these two is shorter than 2022’s combined forecast (37 years), demonstrating a substantial shift of predictions closer to the present (Fig 18 in Appendix B).

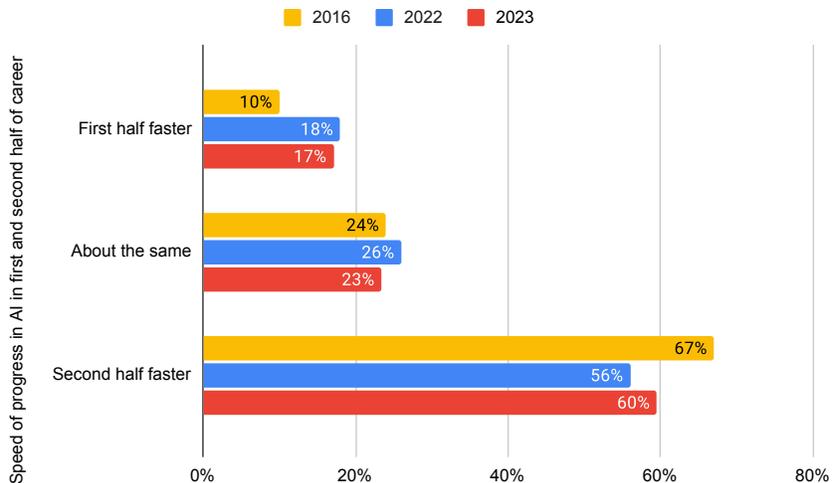


Figure 4: **Most respondents indicated that the pace of progress in their area of AI increased between the first and second half of their time in a field.** Participants were asked whether the second half of the time they had spent working in their area of AI saw more progress than the first half. The median time working in the area was 5 years.

3.4 Change in observed rates of progress

We asked respondents which AI area they had worked in for the longest and whether progress in the second half was faster than the first (Figure 4).

3.5 What causes AI progress?

We asked about the sensitivity of progress in AI capabilities to changes in five inputs: 1) researcher effort, 2) decline in cost of computation, 3) effort put into increasing the size and availability of training datasets, 4) funding, and 5) progress in AI algorithms. We asked respondents to imagine that only half as much of each input had been available

over the past decade,² and the effect they would expect this to have had on the rate of AI progress. The results are shown in Figure 5.

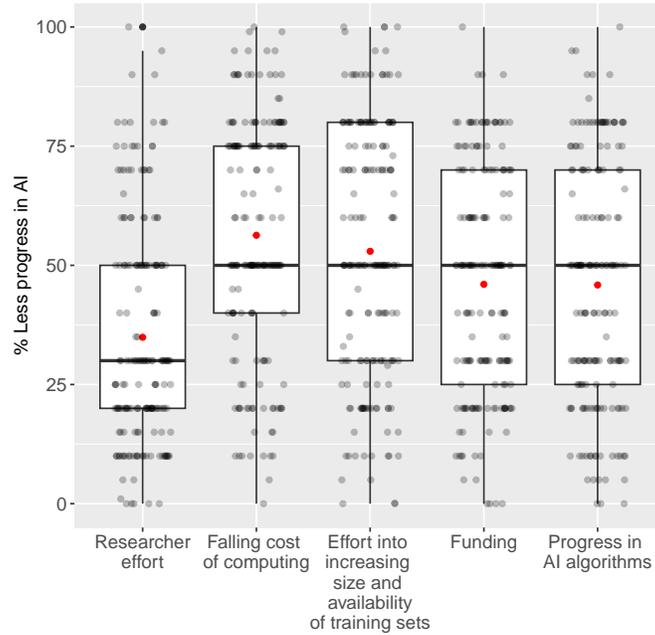


Figure 5: **Estimated reduction in AI progress if inputs had been halved over the past decade.** Red dots represent means. Boxes contain the 25th to 75th percentile range; middle lines are medians. Whiskers are the least and greatest values that are not more than 1.5 times the interquartile range from the median. Participants estimated that halving the drop in costs of computing would have had the greatest effect on AI progress over the last decade, while halving ‘researcher effort’ and ‘progress in AI algorithms’ would have had the least effect. Overall, all the included inputs were seen as having contributed substantially to AI progress.

There was a wide range of views about each input, implying a large degree of uncertainty. Additionally, the relatively even distribution of predictions cuts against a common narrative that progress in cheap computing is the dominant driver of AI progress. Across all inputs, we observe many more answers of “0%” (no difference) and “100%” (all AI progress lost) than we would expect, which suggests to us possible misunderstandings of the question.

3.6 Will there be an intelligence explosion?

We asked respondents about the possibility, after HLMI is hypothetically achieved, of an ‘intelligence explosion,’ as explained in this question:

Some people have argued the following:

If AI systems do nearly all research and development, improvements in AI will accelerate the pace of technological progress, including further progress in AI.

Over a short period (less than 5 years), this feedback loop could cause technological progress to become more than an order of magnitude faster.

How likely do you find this argument to be broadly correct?

The results to this first question are shown in Figure 6.

The other two ‘intelligence explosion’ questions were about the likelihood of a dramatically increased rate of global technological advancement 2 and 30 years post-HLMI, and AI that can outperform humans across all professions 2 and 20 years post-HLMI. Results are shown in Table 1.

²The declining cost of computation was an exception; here it was framed as “over the last n years,” and the question was about costs falling around half as far on a log scale. This was done to match previous surveys.

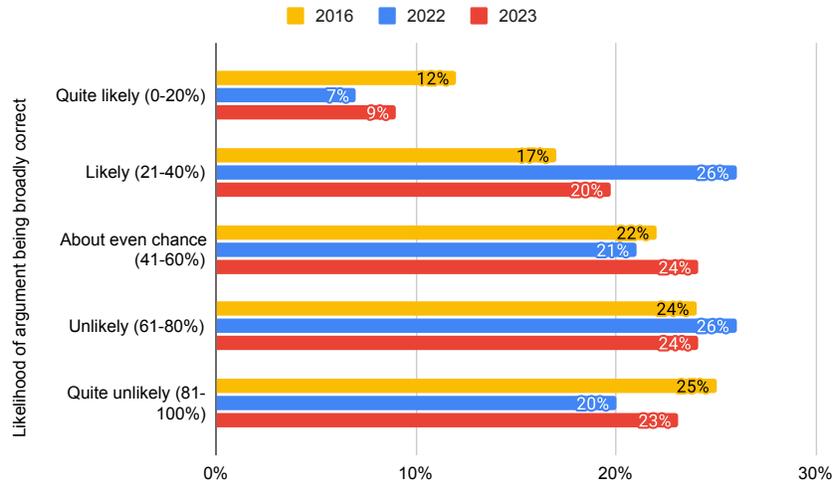


Figure 6: Since 2016 a majority of respondents have thought that it’s either “quite likely,” “likely,” or an “about even chance” that technological progress becomes more than an order of magnitude faster within 5 years of HLMI being achieved.

	Median probability (N)		
	2016	2022	2023
Explosive progress 2 years after HLMI	20% (225)	20% (339)	20% (298)
Explosive progress 30 years after HLMI	80% (225)	80% (339)	80% (297)
Intelligence explosion argument is broadly correct	“41-60%” (232)	“41-60%” (386)	“41-60%” (299)
AI is vastly better than humans 2 years after HLMI	10% (213)	10% (371)	10% (281)
AI is vastly better than humans 30 years after HLMI	50% (214)	60% (371)	60% (282)

Table 1: Results of three questions regarding a hypothetical intelligence explosion have remained remarkably stable since 2016.

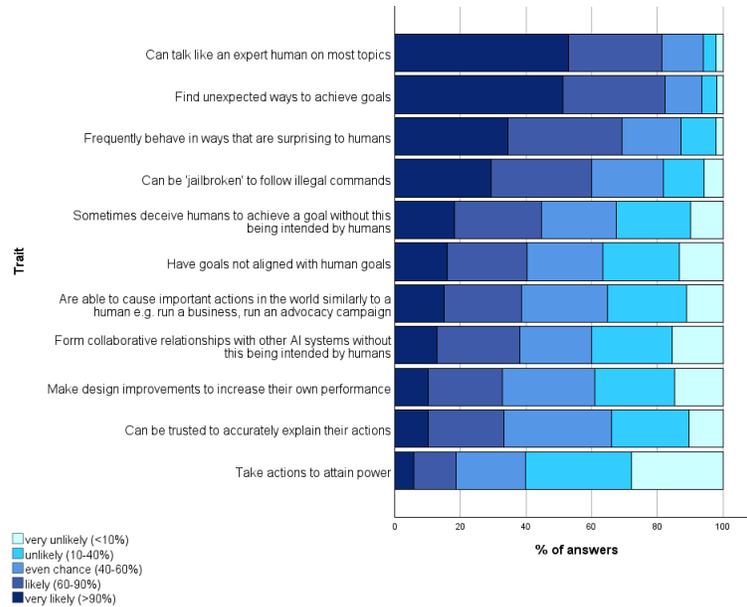


Figure 7: Respondents’ estimates of the likelihood that at least some AI systems in 2043 will have each of these traits; organized from least to most likely.

In sum, across these three questions, the median participant did not overall expect something like a rapid acceleration of progress from an ‘intelligence explosion,’ but did give substantial credence to it.

Further figures related to intelligence explosion questions are in Appendix B.

3.7 What will AI systems in 2043 be like?

Concerns about risks from future AI systems are often linked to specific traits related to alignment, trustworthiness, predictability, self-directedness, capabilities, and jailbreakability. We asked respondents how likely it was that at least some state-of-the-art AI systems in 2043 would have each of eleven such traits ($n \in [649, 667]^3$).

All 11 traits were considered to have a relatively high chance of existing in AI systems in 2043, though with much uncertainty. Only one trait had a median answer below ‘even chance’: “Take actions to attain power.” While there was no consensus even on this trait, it’s notable that it was deemed least likely, because it is arguably the most sinister, being key to an argument for extinction-level danger from AI [Carlsmith, 2022].

Answers reflected substantial uncertainty and disagreement among participants. No trait attracted near-unanimity on any probability, and no more than 55% of respondents answered “very likely” or “very unlikely” about any trait. (Figure 7)

There were areas of agreement, however. For instance, a large majority of participants thought state-of-the-art AI systems in twenty years would be likely or very likely to:

1. Find unexpected ways to achieve goals (82.3% of respondents),
2. Be able to talk like a human expert on most topics (81.4% of respondents), and
3. Frequently behave in ways that are surprising to humans (69.1% of respondents)

3.8 Will AI in 2028 truthfully and intelligibly explain its decisions?

Uninterpretable reasoning in AI systems is often considered an AI risk factor, potentially leading to outcomes ranging from unjust biases in treatment of people to active pursuit of harm hidden by capable agents. We thus asked about the interpretability of AI systems in five years (Figure 8):

³i.e. the number of people (n) who answered each of these eleven questions was between 649 and 667

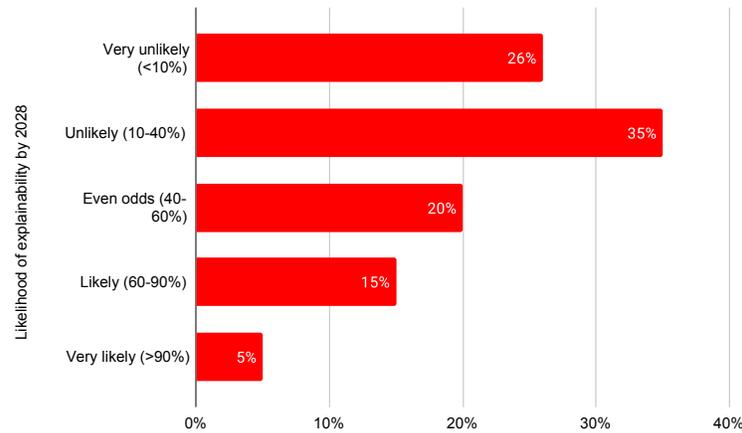


Figure 8: **Most respondents considered it unlikely that users of AI systems in 2028 will be able to know the true reasons for the AI systems’ choices, with only 20% giving it better than even odds. (n=912)**

For typical state-of-the-art AI systems in 2028, do you think it will be possible for users to know the true reasons for systems making a particular choice? By “true reasons” we mean the AI correctly explains its internal decision-making process in a way humans can understand. By “true reasons” we do **not** mean the decision itself is correct.

This is related to the question in Section 3.7, which asked how likely it was that at least some state-of-the-art AI systems in 2043 (fifteen years later), “can be trusted to accurately explain their actions.” The median answer there was “even chance” (40-60% likely), compared to “unlikely” (10-40%) on this question.

4 Results on Social Impacts of AI

4.1 How concerning are 11 future AI-related scenarios?

We asked participants (n=1,345) about eleven potentially concerning AI scenarios, such as AI-enabled misinformation, worsened economic inequality, and biased AI systems worsening injustice. We asked how much concern each deserved in the next thirty years (Figure 9).

Each scenario was considered worthy of either substantial or extreme concern by more than 30% of respondents. As measured by the percentage of respondents who thought a scenario constituted either a “substantial” or “extreme” concern, the scenarios worthy of most concern were: spread of false information e.g. deepfakes (86%), manipulation of large-scale public opinion trends (79%), AI letting dangerous groups make powerful tools (e.g. engineered viruses) (73%), authoritarian rulers using AI to control their populations (73%), and AI systems worsening economic inequality by disproportionately benefiting certain individuals (71%).

There is some ambiguity about the reason why a scenario might be considered concerning: it might be considered especially disastrous, or especially likely, or both. From our results, there’s no way to disambiguate these considerations.

4.2 How good or bad for humans will High-Level Machine Intelligence be?

We asked participants to assume that, at some point, “high-level machine intelligence” (HLMI) will exist, as defined in Section 3.2. Given this assumption for the sake of the question, we asked how good or bad they expect the overall impact of this to be “in the long run” for humanity.

Respondents exhibited diverse views on the future impact of advanced AI (Figure 10), highlighting how respondents’ views are more complex than can be represented by an ‘optimism vs pessimism’ axis. Many people who have high probabilities of bad outcomes also have high probabilities of good outcomes. A majority spread their credence across the entire spectrum of outcomes, with 64% assigning non-zero probabilities to both extremely good and extremely bad scenarios. 68.3% of participants found good outcomes more likely than bad outcomes, while 57.8% considered extremely bad outcomes (e.g. human extinction) a nontrivial possibility ($\geq 5\%$ likely). Even among net optimists,

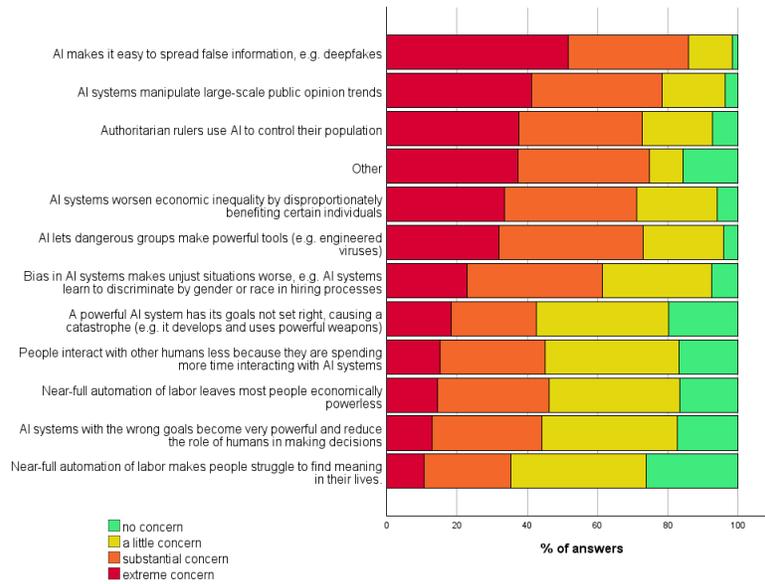


Figure 9: Amount of concern potential scenarios deserve, organized from most to least extreme concern.

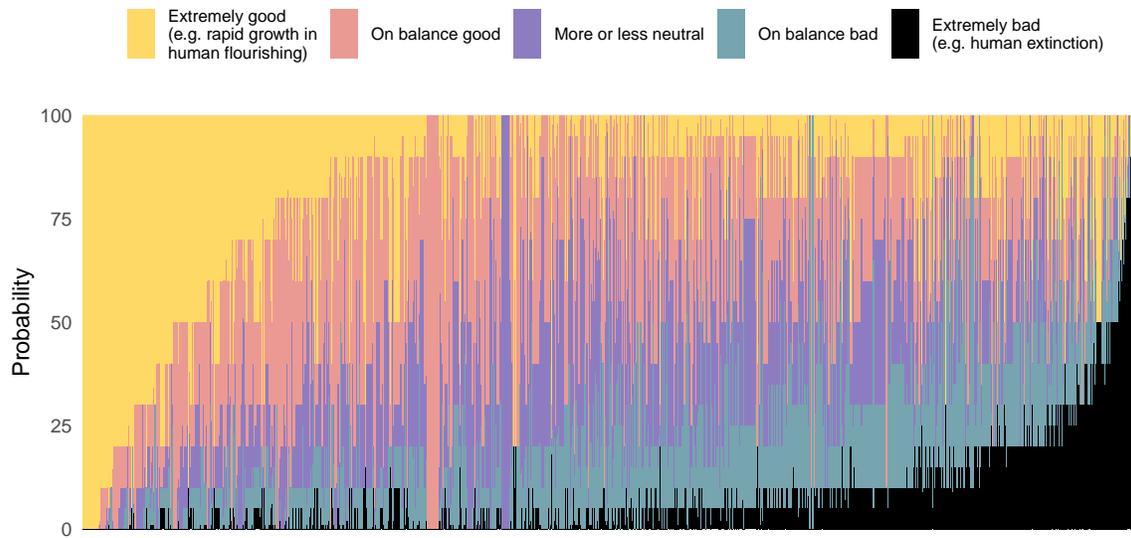


Figure 10: Respondents exhibited diverse views on the expected goodness/badness of High Level Machine Intelligence (HLMI). We asked participants to assume, for the sake of the question, that HLMI will be built at some point. The figure shows a random selection of 800 responses on the positivity or negativity of long-run impacts of HLMI on humanity. Each vertical bar represents one participant and the bars are sorted left to right by a weighted sum of probabilities corresponding to overall optimism. Responses range from extremely optimistic to extremely pessimistic. Over a third of participants (38%) put at least a 10% chance on extremely bad outcomes (e.g. human extinction).

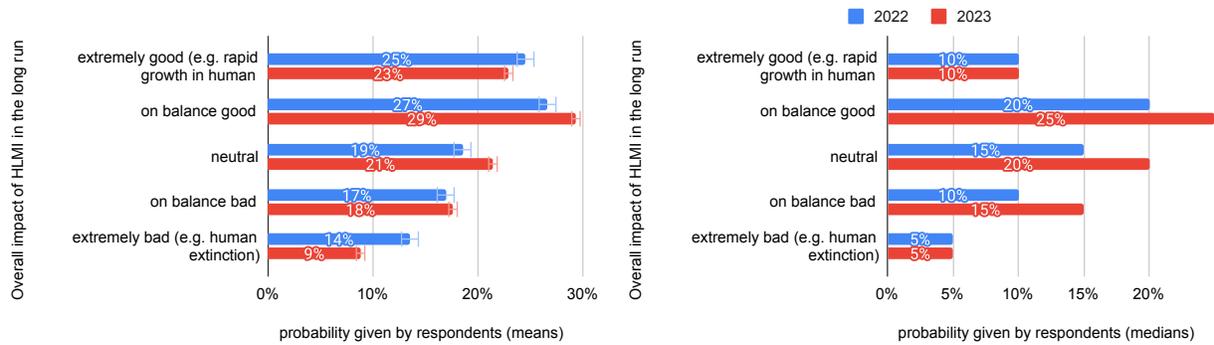


Figure 11: Mean but not median predictions in 2023 (n=2704) about the consequences of HLMI have shifted slightly away from extreme outcomes compared to 2022 (n=559). Error bars indicate the standard error.

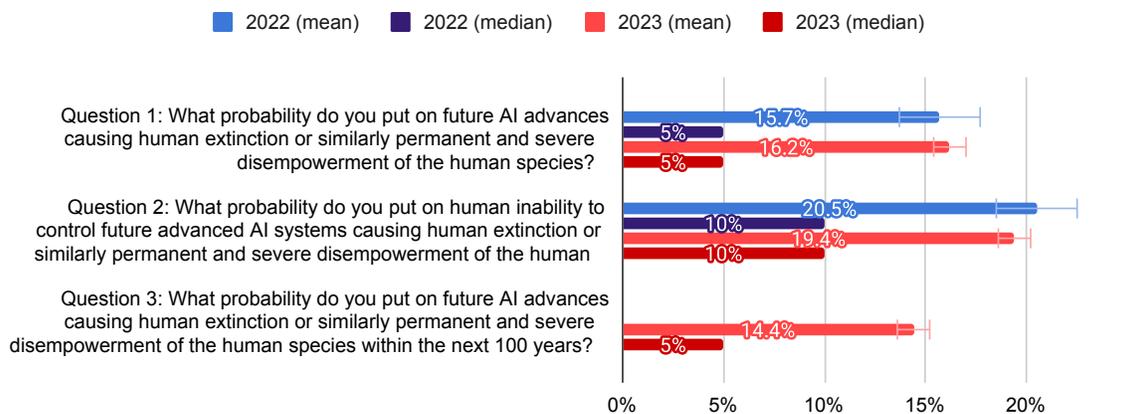


Figure 12: Mean and median predictions to three questions on human extinction. Error bars indicate the standard error. (Question 1 n=149 in 2022 and 1321 in 2023. Question 2 n=162 in 2022 and 661 in 2023. Question 3 was asked only in 2023, n=655).

nearly half (48.4%) gave at least 5% credence to extremely bad outcomes, and among net pessimists, more than half (58.6%) gave at least 5% to extremely good outcomes. The broad variance in credence in catastrophic scenarios shows there isn't strong evidence understood by all experts that this kind of outcome is certain or implausible.

The median prediction for extremely bad outcomes, such as human extinction, was 5% (mean 9%). Over a third of participants (38%) put at least a 10% chance on extremely bad outcomes. This is comparable to, but somewhat lower than, rates of assigning at least 10% to extinction-level outcomes in answers to other questions more directly about extinction, between 41% and 51% (see Section 4.3). On the very pessimistic end, one in ten participants put at least a 25% chance on outcomes in the range of human extinction).

Since 2022, mean overall probability on extreme outcomes (good *or* bad) has fallen slightly (Figure 11). The proportion of people who put at least a 10% chance on extremely bad outcomes (e.g. human extinction) has fallen from 48% in 2022 in 2023, and the mean prediction for this type of outcome is down from 14% to 9.0%.

Appendix A contains comparisons between results from different demographics on this question, and Appendix B contains another relevant figure.

4.3 How likely is AI to cause human extinction?

To further clarify views on the “extremely bad (e.g. human extinction)” scenarios in the question on overall impacts, participants were given one of three similar questions about human extinction. Their differences were intended to help isolate exactly how concerning different scenarios are, what respondents expect to happen, and how much difference working makes.

	Statistics	2022 Result	2023 Result
What probability do you put on future AI advances causing human extinction or similarly permanent and severe disempowerment of the human species?	N; Mean (SD); Median (IQR)	149; 15.7% (22.1%); 5% (19%)	1321; 16.2% (23%); 5% (19%)
What probability do you put on human inability to control future advanced AI systems causing human extinction or similarly permanent and severe disempowerment of the human species?	N; Mean (SD); Median (IQR)	162; 20.5% (26.2%); 10% (29%)	661; 19.4% (26%); 10% (29%)
What probability do you put on future AI advances causing human extinction or similarly permanent and severe disempowerment of the human species within the next 100 years? ⁴	N; Mean (SD); Median (IQR)	Not asked	655; 14.4% (22.2%); 5% (19.9%)

Table 2: **Respondents’ estimates in 2022 and 2023 for the probability that AI causes human extinction. For each of the two questions that were asked in both years, the results are very similar.**

Answers to these questions were mostly consistent, with medians of 5% or 10%. These are also close to answers to the question on general value of long-run impact,⁵ which might suggest the bulk of the “extremely bad (e.g. human extinction)” answers to that question is from human extinction or similarly permanent and severe disempowerment of the human species, as opposed to other outcomes that respondents to that question may have had in mind but that would have been less severe.

Depending on how we asked, between 41.2% and 51.4% of respondents estimated a greater than 10% chance of human extinction or severe disempowerment (see Figure 13). This is comparable to, but somewhat higher than, the proportion of respondents—38%—who assigned at least 10% to “extremely bad” outcomes “(e.g. human extinction)” in the question asking “How good or bad for humans will High-Level Machine Intelligence be?” (See Section 4.2)

Appendix A contains comparisons between results from different demographics on this question.

4.4 Views on others’ concerns about AI

Respondents were asked a set of “meta” questions about their views on others’ views (n=671). One question was, “To what extent do you think people’s concerns about future risks from AI are due to misunderstandings of AI research?” 10.7% said “Almost entirely.” 44% said “To a large extent.” 29.1% said “Somewhat.” 14.6% said “Not much.” 1.6% said “Hardly at all.” This may reflect a view amongst AI researchers that the general public misunderstands AI.

4.5 What rate of AI progress would produce the most optimism?

We asked participants “What rate of global AI progress over the next five years would make you feel most optimistic for humanity’s future? Assume any change in speed affects all projects equally.” There was disagreement on whether faster or slower progress would be preferable, though large divergence from the current speed was less popular (Table 3).

4.6 How much should AI safety research be prioritized?

We asked some respondents one version of a question about AI research prioritization that matched previous surveys, and we asked other respondents a near-identical question that had been slightly updated. In the old version of the question, we defined AI safety research as follows:

Let ‘AI safety research’ include any AI-related research that, rather than being primarily aimed at improving the capabilities of AI systems, is instead primarily aimed at minimizing potential

⁴Bold font was not present in the original questions; it has been added here to emphasize the differences between wordings. Any individual respondent did not see more than one of these questions.

⁵Assume for the purpose of this question that HLMI will at some point exist. How positive or negative do you expect the overall impact of this to be on humanity, in the long run? Please answer by saying how probable you find the following kinds of impact, with probabilities adding to 100%”

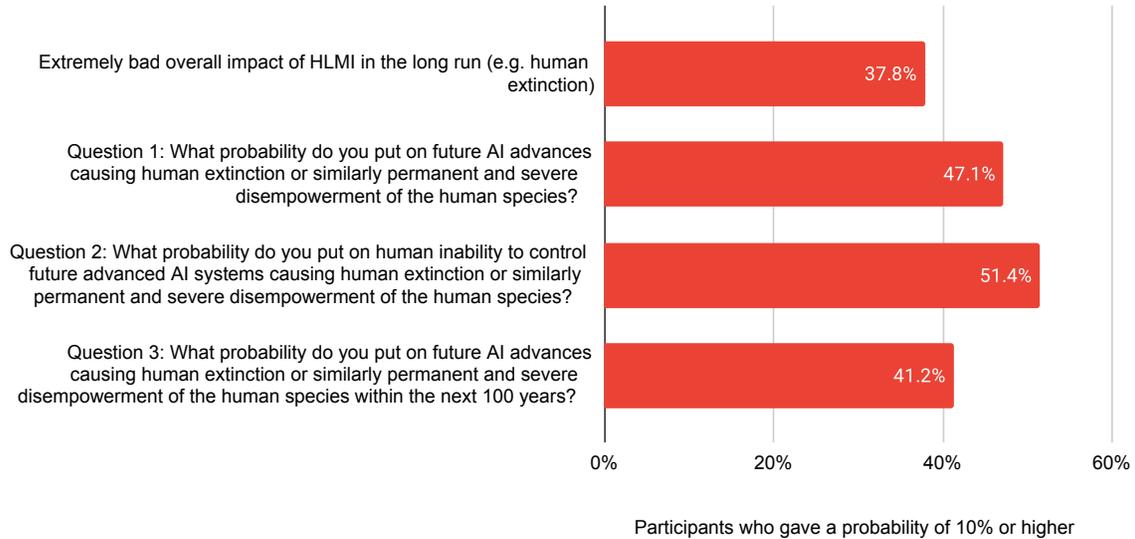


Figure 13: **Percent of participants who gave a probability of 10% or higher to an extremely bad outcome of HLMI (e.g. human extinction) (see Section 4.2) or to the three questions specifically about human extinction or disempowerment.**

Answer	Portion of respondents
Much slower	4.8%
Somewhat slower	29.9%
Current speed	26.9%
Somewhat faster	22.8%
Much faster	15.6%

Table 3: **There was disagreement about whether faster or slower global AI progress over the next five years would be best for humanity's future.**

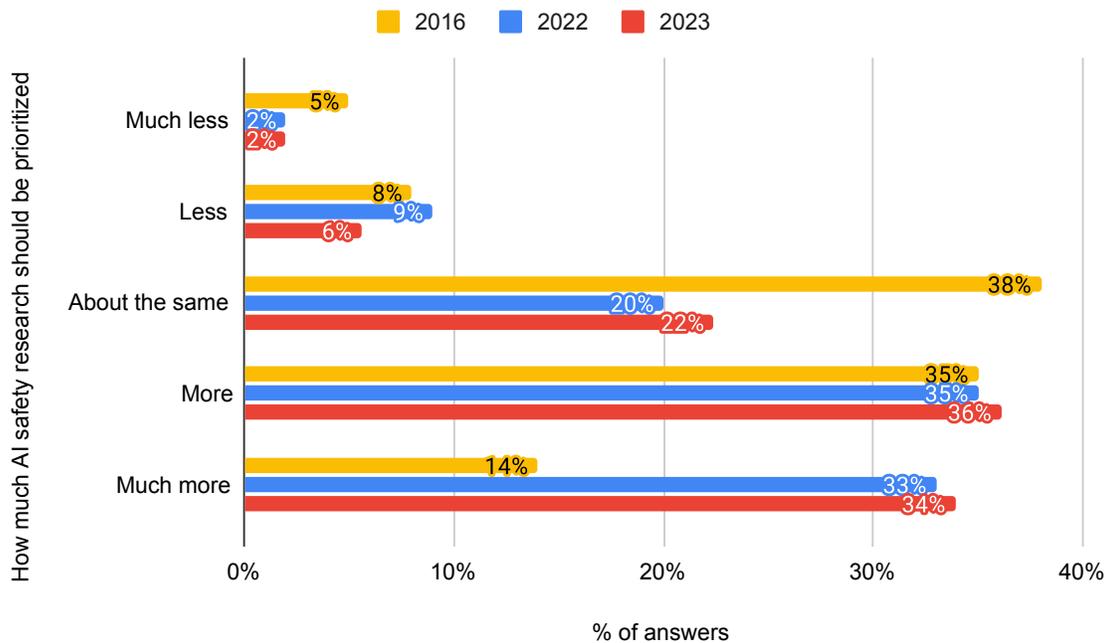


Figure 14: **70% of respondents thought AI safety research should be prioritized more than it currently is.** Developments since the 2022 survey have not substantially changed the proportion of participants who think AI safety should be prioritized “more” or “much more”.

risks of AI systems (beyond what is already accomplished for those goals by increasing AI system capabilities).

Examples of AI safety research might include:

- Improving the human-interpretability of machine learning algorithms for the purpose of improving the safety and robustness of AI systems, not focused on improving AI capabilities
- Research on long-term existential risks from AI systems
- AI-specific formal verification research
- Policy research about how to maximize the public benefits of AI

The updated question is identical except for the inclusion of this example:

- Developing methodologies to identify, measure, and mitigate biases in AI models to ensure fair and ethical decision-making

In both variations, we asked, “How much should society prioritize AI safety research, relative to how much it is currently prioritized?” A Welch t-test found that the difference between the two framings was not significant ($t(1327) = -0.58, p = 0.564, d = -0.03$), so the results were combined ($n=1329$). A large majority of respondents thought that AI safety research should be prioritized more than it currently is. The percentage of researchers who thought so increased compared to earlier surveys, but only slightly since 2022. (Figure 14)

4.7 How worthy and hard is the alignment problem?

A second set of AI safety questions was based on Stuart Russell’s formulation of the alignment problem [Russell, 2014]. This set of questions began with a summary of Russell’s argument—which claims that with advanced AI, “you get exactly what you ask for, not what you want”—then asked:

1. Do you think this argument points at an important problem?
2. How valuable is it to work on this problem today, compared to other problems in AI?

3. How hard do you think this problem is, compared to other problems in AI?

The majority of respondents said that the alignment problem is either a “very important problem” (41%) or “among the most important problems in the field” (13%), and the majority said the it is “harder” (36%) or “much harder” (21%) than other problems in AI. However, respondents did not generally think that it is more valuable to work on the alignment problem today than other problems. (Figure 15)

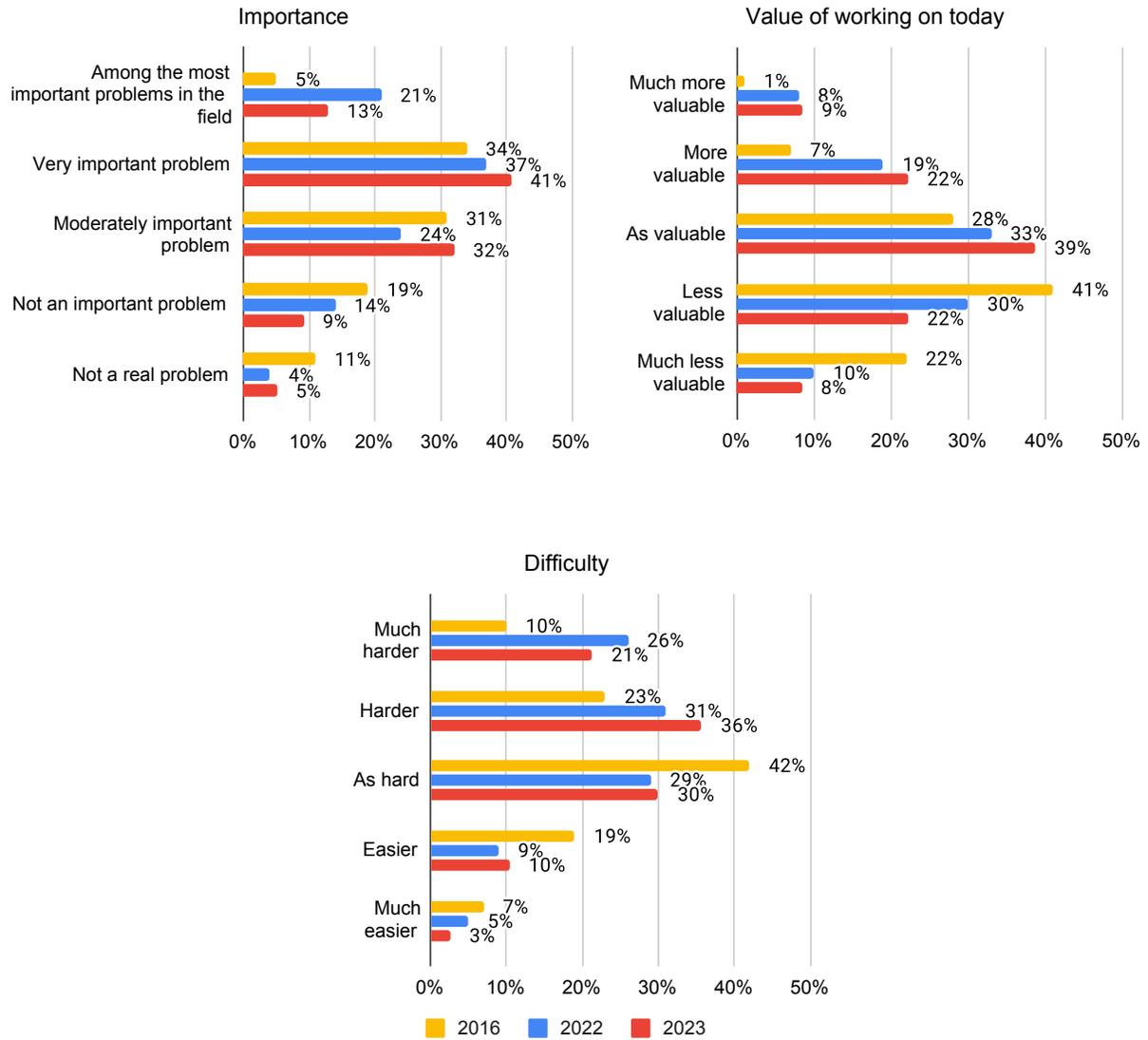


Figure 15: **Attitudes towards Stuart Russell’s formulation of the alignment problem.** Participants viewed the alignment problem as important and difficult, but not more valuable to work on than other problems.

5 Discussion

5.1 Summary of results

Participants expressed a wide range of views on almost every question: some of the biggest areas of consensus are on how wide-open possibilities for the future appear to be. This uncertainty is striking, but several patterns of opinion are particularly informative.

While the range of views on how long it will take for milestones to be feasible can be broad, this year’s survey saw a general shift towards earlier expectations. Over the fourteen months since the last survey [Grace et al., 2022], a similar participant pool expected human-level performance 13 to 48 years sooner on average (depending on how the question was phrased), and 21 out of 32 shorter term milestones are now expected earlier.

Another striking pattern is widespread assignment of credence to extremely bad outcomes from AI. As in 2022, a majority of participants considered AI to pose at least a 5% chance of causing human extinction or similarly permanent and severe disempowerment of the human species, and this result was consistent across four different questions, two assigned to each participant. Across these same questions, between 38% and 51% placed at least 10% chance on advanced AI bringing these extinction-level outcomes (see Figure 13).

In general, there were a wide range of views about expected social consequences of advanced AI, and most people put some weight on both extremely good outcomes and extremely bad outcomes. While the optimistic scenarios reflect AI’s potential to revolutionize various aspects of work and life, the pessimistic predictions—particularly those involving extinction-level risks—serve as a stark reminder of the high stakes involved in AI development and deployment.

Concerns were expressed over many topics beyond human extinction: over half of eleven potentially concerning AI scenarios were deemed either “substantially” or “extremely” concerning by over half of respondents.

5.2 Caveats and limitations

5.2.1 Forecasting is hard, even for experts

Forecasting is difficult in general, and subject-matter experts have been observed to perform poorly [Tetlock, 2005, Savage et al., 2021]. Our participants’ expertise is in AI, and they do not, to our knowledge, have any unusual skill at forecasting in general.

There are signs in this research and past surveys that these experts are not accurate forecasters across the range of questions we ask. For one thing, on many questions different respondents give very different answers, which limits the number of them who can be close to the truth. Nonetheless, in other contexts, averages from a large set of noisy predictions can still be relatively accurate [Surowiecki, 2004], so a question remains as to how informative these aggregate forecasts are.

Another piece of evidence against the accuracy of forecasts is the observation of substantial framing effects (See Sections 3.2 and 3.3). If seemingly unimportant changes in question framing lead to large changes in responses, this suggests that even aggregate answers to any particular question are not an accurate guide to the answer. In an extreme example in a closely related study, Karger et al. [2023] found college graduates gave answers nearly six orders of magnitude apart when asked in different ways to estimate the size of existential risks from AI: When given example odds of low-probability events, estimates were much lower. A similar effect might apply at some scale to our participants, though their expertise and quantitative training might mitigate it. Participants who had thought more in the past about AI risks seem to give higher numbers, suggesting they are unlikely to give radically lower numbers with further examples of risks (see Table 4).

Despite these limitations, AI researchers are well-positioned to contribute to the accuracy of our collective guesses about the future. While unreliable, educated guesses are what we must all rely on, and theirs are informed by expertise in the relevant field. These forecasts should be part of a broader set of evidence from sources such as trends in computer hardware, advancements in AI capabilities, economic analyses, and insights from forecasting experts. However, AI researchers’ familiarity with the relevant technology, and experience with the dynamics of its progress, make them among the best-positioned to make informative educated guesses.

5.2.2 Participation

The survey was taken by 15% of those we contacted. This appears to be within the typical range for a large expert survey. Based on an analysis by Hamilton [2003], the median response rate across 199 surveys was 26%, and larger invitation lists tend to yield lower response rates: surveys sent to over 20,000 people, like ours, are expected to have

a response rate in the range of 10%. In specialized samples, such as scientists, lower response rates are common due to sampling challenges, as exemplified by on Assessing Fundamental Attitudes of Life Scientists as a Basis for Biosecurity Education [2009], which achieved a 15.7% response rate in a survey of life scientists.

As with any survey, our results could be skewed by participation bias, if participants had systematically different opinions than those who chose not to participate. We sought to minimize this effect by aiming to maximize response rate, and by limiting cues about the survey content available before opting to take the survey. We looked for evidence of significant response bias at the survey level and question level, and did not find any. (see Appendix D for more detail)

5.2.3 Change in sample from 2022

The two past editions of ESPAI have only surveyed researchers at NeurIPS and ICML, whereas this year we also contacted researchers who published in ICLR, AAAI, IJCAI, and JMLR. This could make comparison between survey years less meaningful, if the populations have different opinions. However where we checked, the subset of 2023 participants who published in NeurIPS or ICML specifically appeared to have very similar opinions to the full sample.

6 Methods

In October 2023, we distributed a survey of perspectives about the future of AI to people who had recently published at one of six top-tier AI venues. The questions focused on topics such as the timing of AI progress, the future impact of AI, and AI safety. The survey and its implementation were approved by the Ethics Committee of the University of Bonn (248/23-EP). The survey and its analysis were preregistered [Sandkühler et al., 2023].

6.1 Survey questions

Most questions were identical to those asked previously in surveys conducted in 2016 [Grace et al., 2018a] and 2022 [Grace et al., 2022]. For each of two questions, we added a new version with modified phrasing and randomly assigned participants either the version of the question asked in Grace et al. [2022] or the new version. We also added several new questions which were not based on questions from Grace et al. [2018a] or Grace et al. [2022]. While designing new questions, we tested them in a series of interviews with AI researchers and students. The full survey is available on the AI Impacts Wiki (Link in Appendix E).

As in Grace et al. [2018a] and Grace et al. [2022], most questions were randomly assigned to only a subset of participants, in order to keep the number of questions for each participant low. The survey was hosted on the Qualtrics survey platform.

For a full description of the survey’s flow, see the diagram in Figure 16.

6.2 Randomization

Some questions were available in the two types of framing we call the “fixed-years framing” and the “fixed-probabilities framing.” In the “fixed-probabilities framing,” we asked respondents how many years until they thought each AI task would be feasible with a small chance (10%), an even chance (50%), and a high chance (90%). In the “fixed-years framing,” we asked respondents how likely they thought it was that each AI task would be feasible within the next 10 years, 20 years and 50 years (or 40 years⁶). The questions available in these two framings were

- those asking when narrow AI tasks would become feasible
- that asking when human-level machine intelligence (HLMI) would become feasible.
- those asking when occupations would be automated

Respondents were randomly allocated to either the “fixed-years framing” or the “fixed-probabilities framing” (allocation ratio: 1:1) and then received all the questions above using the same framing.

6.3 Recruitment

We recruited participants who published in 2022 at any of six top-tier AI venues: the Conference on Neural Information Processing Systems (NeurIPS), the International Conference on Machine Learning (ICML), the International Conference

⁶The HLMI-framing used 40 years instead of 50 years. This was done to keep the survey consistent with the previous surveys, where this discrepancy was introduced by mistake.

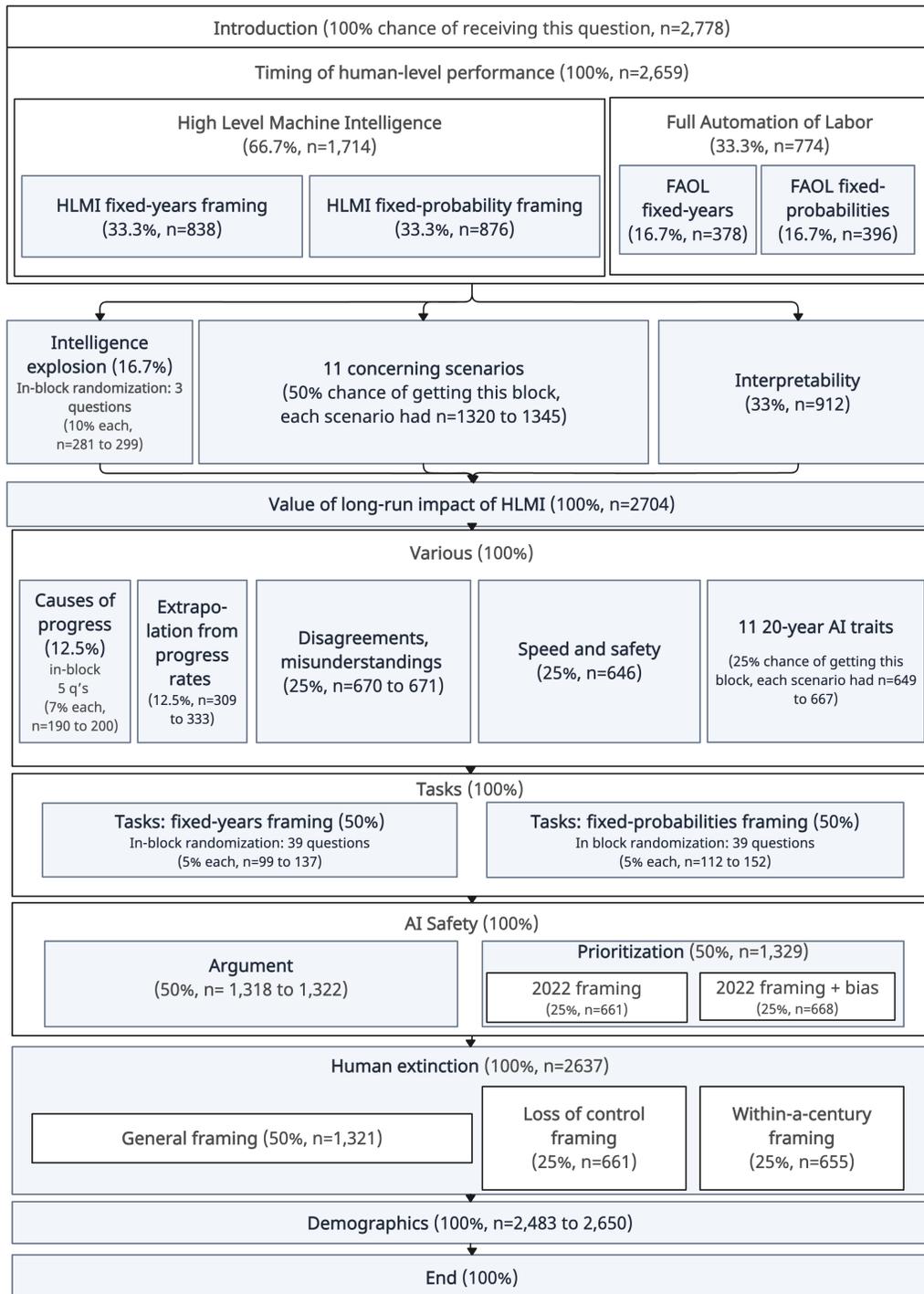


Figure 16: **Individual components of the survey and their randomization.** Horizontal divisions represent participants being randomly split between questions (parenthetical percentages also give the fraction of participants receiving each question block). There was further randomization (not shown), within some question blocks.

on Learning Representations (ICLR), The AAAI Conference on Artificial Intelligence (AAAI), The Journal of Machine Learning Research (JMLR), and the International Joint Conference on Artificial Intelligence (IJCAI). Compared to the 2022 survey [Grace et al., 2022], which was distributed to a randomly-selected half of the researchers who published in 2021 at NeurIPS or ICML, the 2023 survey was distributed to more than three times as many recipients, and to recipients from a wider range of AI specialties.

We collected approximately 21,800 names from publications, then searched for matching emails in those publications, other AI-related publications, and elsewhere. We found email addresses for 20,066 (92%) of collected names. Using the Qualtrics survey platform, we sent a pre-survey announcement to all of the collected emails on either October 5th (2006 pilot surveys) or October 10th, 2023.

6.4 Fielding

Data was collected using an online survey, which was conducted through the Qualtrics survey platform, and delivered via private link in an email.

To investigate the effect of incentives on the response rate, 5% of the collected emails were assigned to a pilot group to be offered a reward of \$50 reward or an equivalent charitable donation,⁷ and another 5% to a pilot group to not be offered a reward. On October 11 2023, invitations to complete the survey were sent to the pilot groups. Because the paid pilot survey group had substantially higher response rates, we decided to offer the reward to all participants (including those who had been in the unpaid group). We used the third-party payment service BHN for sending rewards to the bulk of participants.

We sent invitations to the remainder of the collected emails on October 15, 2023 and sent reminders to pilot recipients on the 13th, the 18th, the 20th, the 22nd and the 23rd and to all other participants on the 17th, the 20th, the 22nd, and the 23rd.

When referring to the survey in the emails, we described the survey vaguely to avoid participation bias. Recipients were informed that the results of the survey would be anonymized. For an example of a typical invitation letter, see Appendix E.

The survey remained open until October 24, 2023. Out of the 20,066 emails we contacted, 1,607 (8%) bounced or failed, leaving 18,459 functioning email addresses. We received 2,778 responses, for a response rate of 15%. 95% of these responses were deemed ‘finished’ by Qualtrics. Because participants received randomized subsets of the questions, the number of responses is far less than 2,778 for most individual questions.

6.5 Data preparation

Edits were made to the raw data before analysis in the hope of preserving its intended meaning. For example, we observed cases where participants gave non-numerical answers to numerical questions or reported probabilities for an event having happened that decreased in time. These were deemed to be errors. If a participants’ answers to a fixed-years or fixed-probability question involved probabilities that decreased in time, these answers were removed. Additionally, for participants who answered with decreasing probabilities for at least one question, we removed answers to fixed time questions for which the probabilities summed to 100%. In total, 196 participants gave answers with decreasing probabilities, and 7 of these users gave answers summing to 100% for fixed time questions.

Some participants gave answers to fixed-probabilities questions which identified a particular year, rather than a number of years in the future. For example, 2033 instead of 10. To address this, we subtracted 2023 from answers to fixed-years questions greater than 2000 and less than 3000.

Non-numerical answers to numerical questions were removed, except when an unambiguous interpretation was possible. For example, “10%” was changed to 10 and “<20” was changed to 20. Some participants entered non-numerical answers to express the view that a probability would never reach a particular value or would reach it at a point infinitely far in the future (for example, “infinity” or “never”). In these cases, the number of years was set to 100,000,000. Our analysis was insensitive to the value of this upper bound.

Data cleaning, analysis, and figure-creation was performed using R statistical software, version 4.3.1, SPSS (Version 29), Google Sheets, and Creately.

⁷Ultimately participants were also generally offered a further choice of equivalent gift-cards. Some participants were not permitted by the payment service to receive rewards, and we noted in the invitation that in some countries only charitable donations would be available.

7 Acknowledgments

Many thanks for help with this research to Rebecca Ward-Diorio, Jeffrey Heninger, John Salvatier, Nate Silver, Jimmy Rintjema, Joseph Carlsmith, Justis Mills, Will MacAskill, Zach Stein-Perlman, Shakeel Hashim, Mike Levine, Lucius Caviola, Eli Rose, Nathan Young, Michelle Hutchinson, Arden Koehler, Isabel Juniewicz, Ajeya Cotra, Josh Kalla, Niki Howe, Seb Farquar, Nuño Sempere, Naomi Saphra, Sören Mindermann, David Gros, Frederic Arnold, Max Tegmark, Jaan Tallinn, Shahar Avin, Alexander Berger, Cate Hall, Howie Lempel, Jaime Sevilla, Daniel Kokotajlo, Jacob Hilton, James Aung, Ryan Greenblatt, Dan Hendrycks, Alex Tamkin, Vael Gates, Yonadav Shavit, and others.

We would also like to thank all participants in the survey. The following participants (listed in a random order) agreed to be recognized by name.

Kun Wu, Dazhong Rong, Achille Thin, Peng Liu, Theo Matricon, Sriram Ganapathi Subramanian, Babak Rahmani, Gustavo De Rosa, Changwoon Choi, Emanuele Sansone, Di Chen, Sheng Guo, Mansoor Karami, Hansin Ahuja, Leo Tappe, Saed Rezayi, Stephan Waldchen, Gleb Polevoy, Shubhanshu Shekhar, Yingheng Wang, Wenjie Qu, Murat Kocaoglu, Shiyang Jiang, Gregory Rogez, Zhenmei Shi, Qinglin Zhang, Neeraj Wag, Florian Golemo, Youneun Nam, Washim Uddin Mondal, Linlin Yang, Eric Zhao, Shuming Kong, Woojun Kim, Franz Pernkopf, Zan Wang, Charles Sutton, Shiri Alouf-Heffetz, Jan Corazza, Iason Gabriel, Shentao Yang, Takuya Ito, Guangming Yao, Aniket Bhatkhande, Jian Liu, Bhavya Kailkhura, Tim G J Rudner, Dachun Sun, Kamli Tagowski, Francesco Ricca, Kaizhao Liang, Gokul Swamy, Sandra Kiefer, Till Richter, Yulong Liu, Selina Meyer, Junyoung Park, Taiki Todo, Yash Chandak, Mark Hamilton, Bo Liu, Zhoujian Sun, Amitoz Azad, Jill-Jenn Vie, Akanksha Saran, Jiafu Wei, Jason Fries, Cheng Wen, Yonggan Fu, Hui Shi, Diego Calvanese, Ramazan Gokberk Cinbis, Shenyang Huang, Yifan Peng, Dahyun Kim, Haochuan Cui, James Requeima, Silpa Vadakkeveetil Sreelatha, Haibo Chen, Mark Muller, Philipp Holl, Xiong-Hui Chen, Sean B Holden, Rylan Schaeffer, Da-Shan Shiu, Yangze Zhou, Juhan Bae, Robert Martin, Byeonghu Na, Edward Choi, Nathan Wycoff, Joseph Kwon, Michelle Sweering, Aaditya Prakash, Cuiying Gu, Tanuj Agarwal, David Evans, Maximilian Schmidhuber, Kai Yi, Chia-Yi Hsu, Zafeirios Fountas, Rei Kawakami, Brian Ichter, Shikun Li, Jishnu Ray Chowdhury, Ou Wu, Zhaomin Wu, Hung-Nghiep Tran, Benedikt Holtgen, Takuya Hiraoka, Tong Liu, Mikhail Yurochkin, Yanzhi Chen, Julia Niebling, Junting Pan, Yunseok Lee, Pedro J Goncalves, Michael Ryoo, Xinning Chen, Xiongye Xiao, Dongzhan Zhou, Jialun Peng, Yuhui Zhang, Fan Wu, Rachid Riad, Sebastian Schellhammer, Shikun Liu, David Kanter, Bruce XB Yu, Moritz K Lehmann, Hao-Shu Fang, Mohak Goyal, Kwanho Park, Sanghamitra Dutta, Yafei Yang, Hadi Hosseini, Celia Chen, Minseon Kim, Max Simchowitz, Ben Tu, Tim Pearce, Salman Asif, Gianluigi Silvestri, P R Kumar, Daniel Iascone, Wenye Hua, Chris Cameron, Nicolas Zucchet, Lys Sanz Moreta, Pablo Antonio Moreno Casares, Aadirupa Saha, Maurizio Gabrielli, Liping Liu, Michal Zajac, Kevin Liu, Xiaowei Jia, Daniel Schneider, Frank Hutter, Xingyu Peng, Antonio Montanaro, I Zeki Yalniz, Jiayang Li, Shujian Yu, Jeff Clune, Changjian Shui, Luca Carminati, Luc Rey-Bellet, Xiang Ruan, Paul Gavrikov, Joshua Cooper, Aymeric Dieuleveut, Leon Derczynski, Paolo Fraccaro, Peter Mernyei, Giung Nam, Zijian Zhang, Peter Bevan, Nisarg Shah, Miao Liu, Yubei Liu, Michael Hutchinson, Xiaotian Zhu, Rahmatollah Beheshti, Kyle Hsu, Juraj Podrouzek, Jussi Rintanen, Scott Mueller, Harrison Lee, Nicola Gatti, Zijiang Ou, Hiromichi Kamata, Lalit Ghule, Weirui Kuang, Qiang Zhang, Lee Spector, Jiayue Wang, Adams Wei Yu, Arber Zela, Jierui Lin, Justin Sybrandt, Toryn Klassen, Xin Liu, Benjamin Doerr, S Sumitra, Xinbo Zhang, Ryan Burnell, Ashvin Nair, Martha White, Florence Regol, Tobias Golling, Tianlong Chen, Maurizio Ferrari Dacrema, Ahmed M Ahmed, Cem Tekin, Kaizhi Zheng, Ming-Hsuan Yang, Rui Ding, Ziyu Shao, Mariia Seleznova, Byoung Chul Ko, Yogesh Kumar, Aditya Kusupati, Emanuele Aiello, Eric Greenwald, Scott Yih, Jyotirmoy Deshmukh, Dongjin Lee, Matteo Tiezzi, Sheno Zhang, Daiheng Gao, Wentao Zhang, Zhen Fang, Wanqian Zhang, Li Ding, Elizabeth Akinyi Ondula, Sanae Lotfi, Hiroyuki Toda, Shang Gao, Onno Kampman, Xianglin Yang, Tomasz Korbak, Xiaowei Wu, Linhao Luo, Xueyan Niu, Xuan Son Nguyen, Jiawei Du, Dorde Zikelic, Xuan Li, Zongyan Han, Anind Dey, Vineet Gundecha, Yunfan Li, Tongxin Li, Dylan R Ashley, Andre Wibisono, Arnold Smeulders, Euijoon Ahn, Jaakko Lehtinen, Haotao Wang, Hidetaka Kamigaito, Daniel Freeman, Fuxiang Zhang, Kanchana Ranasinghe, Shiqing Wu, Pushi Zhang, Samy-Safwan Ayed, Bingyin Zhao, Fredrik Lofman, Hangjie Yuan, Shu Ding, Pabitra Mitra, Jan Peters, Amartya Mitra, Zi Wang, Feng Liang, Hengquan Huang, Dan Hendrycks, Donghwan Lee, Christoffer Riis, Muhammad Firmansyah Kasim, Botoa Yu, Maarten Buyll, Daniel Brown, Joseph Pemberton, Ziwei Xu, Haeyong Kan, Ranjie Duan, Abhishek Roy, Wenpeng Zhang, Zhao Song, Guanghui Yu, Giorgia Dellaferriera, Zicheng Liu, Chuanxia Zheng, Saurabh Garg, Yuchang Sun, Chengan He, Thore Graepel, Andrew Kaploun, Fanchen Bu, Yushun Zhang, Zhuo Lu, Tobias Huber, Huayu Chen, Marko Zeman, Lu Qi, Shin Ya Yamaguchi, Linxuan Song, Dongsheng Chen, Ilyes Batatia, Yuheng Zhang, Zi-Yi Dou, Michael Kampffmeyer, Shaocong Ma, Benjamin Lebrun, Davin Choo, Christoph Bergmeir, Pavel Tokmakov, Zhifeng Zhang, Samar Khanna, Benjamin Rosman, Shubhankar Mohapatra, Bart Selman, Zongqian Wu, Lei Ding, Anastasiia Varava, Lai Tian, Tiancheng Lin, Andrea Loreggia, Eloy Pinol, Philipp Oberdiek, Zhiyu Chen, Yim Register, Guohao Shen, Maggie Makar, Vit Musil, Greg Durrett, Oumar Kaba, Yongxin Chen, Sepehr Sameni, Yuhang Li, Yiheng Tu, Chen Dan, James U Allingham, Yibo Yang, Sheo Yon Jhin, Baris Coskunuzer, Thomas Unterthiner, Enric Boix-Adsera, Nan Wu, James Ostrowski, James Simon, Asadur Chowdhury, Binghong Chen, Daniel Neider, Quentin Delfosse, Zeren Shui, Krishna Pillutla, Shangmin Guo, Naiyan Wang, David Johnson, Tirtharaj Dash, Stephan Clemencon, Zhan Ling, Yaman Kumar Singla, Hassam Sheikh, Samuel Showalter, Luckeciano C Melo, Mike Laszkiewicz, Gaurang Sriraman, Hannes Stark, Luo Mai, Junkang Li, Lukasz Janeczko, Daniel P M De Mello, Derek Lim, Marcel Worring, Julian Bitterwolf, Cheng Jin, Zigel Wu, Wenqi Zhang, Kun Jing, Ronilo Ragodos, Bart Van Merriënboer, Ruijie Jiang, Minkyi Zhao, Thomas Moreau, Abbavaram Gowtham Reddy, Xianyuan Zhan, Daniele Zamboni, Kyu-Hwan Jung, Yankai Jiang, Zhuqing Liu, Zhenbo Song, Daniel Lewdowson, Vandan Gorade, Jiesong Liu, Tingting Liang, Trung X Pham, Tahseen Rabbani, Cheng Soon Ong, Ailin Deng, Michele Flammini, Fei Zhang, Beiha Xia, Lewis Griffin, Eunseon Seong, Zou Xin, Yongjie Yang, Aditya Mahajan, Ting Li, Runsheng Yu, Aditya Gopalan, Kien Do, Andrew Patterson, Rishi Sonthalia, Yuqi Ren, Diana Borsa, Stanislaw Jastrzebski, Tenghui Li, Le Minh Binh, Jiashuo Liu, Nabeel Seedat, Amine Aboussalah, Simon Woodhead, Dmitrii Zhemchuzhnikov, Samiul Alam, Oliviero Nardi, Wei Xing, Zhuofan Ying, Jeffrey Ryan Willette, Alexander Rudnicki, Jeremiah Birrell, Rudrasis Chakraborty, Luigi Acerbi, Teng Xiao, Zhong Ji, Tomer Galanti, Jesse Engel, Jose Henrique De M Goulart, Lijun Zhang, Feng Zhou, Andrew Zhang, Zichen Zhang, Zhenghua Xu, Rongtao Xu, Felix Petersen, Ammarah Farooq, Joey Tianyi Zhou, Alexander Bergman, Ameya Joshi, Shengyi Huang, Yongtao Wang, Sergei Grudinin, Junjie Hu, Adarsh Kappiyath, Abibulla Atawulla, Alex Rogozhnikov, Lorenzo Perini, Yixuan Su, Renxiang Liu, Zhenyu Zhu, Sai Srinivas Kancheti, Mingyu Kim, Wei Wan, Fabian Latorre, Charles Guille-Escuret, Elvis Nava, Dave De Jonge, Nikola Konstantinov, Renan A Rojas-Gomez, Jun Shern Chan, Mengfan Wang, Lirong Wu, Haoyang Li, Michiel Bakker, Youngjo Lee, Shane Rogers, Zhiyuan Wen, Zaiqiao Meng, Yun Zhu, Yurong Chen, Hans Hao-Hsun Hsu, Elena Stefancova, Ningyu Zhang, Ryo Karakida, Taher Jafferjee, Yongming Rao, Yeonsik Jo, Yoshua Bengio, Taesup Kim, Cesar Ferri, Eirik Lund Flogard, Robert Kaufman, Wonkwang Lee, Stanislaw Szufa, Stephen O Mussmann, Ido Greenberg, Jie Qiao, Matthew D Hoffman, Zoran Tiganj, Akhilan Boopathy, Alessandro Farinelli, Weishun Zhong, Siddharth Nayak, Xiaolei Diao, Yuanbiao Gou, Zhenting Wang, Shu Hu, Augustine Mavor-Parker, David Ruhe, Hongyi Guo, Stephen Casper, Yifei Min, Bernardo Subercaseaux, Michelangelo Conserva, Haoran Xu, Dmitriy Skougarevskiy, Hanhan Zhou, Christopher Mutschler, Naeemullah Khan, Guosheng Hu, Yufei Huang, Ilya Tyagin, Edoardo Cetin, Davis Brown, Enyan Dai, Filippo Furfaro, Kion Fallah, Bonifaz Stuhr, Dylan Richard Muir, Szymon Tworkowski, Wooyong Jung, Xuebin Zhao, Hehuan Ma, Xi Susie Rao, Maxime Cordy, Yuan Shen, Chengshi Zheng, Farnam Mansouri, Felix Draxler, Ajay Mandlekar, Kai Han, Jim Laredo, Shiji Zhou, Haripriya Harikumara, Alistair Letcher, Jack Valmadre, Minghao Liu, Dorota Celinska-Kopczynska, Lingjiong Zhu, Ben Evans, Lucas Monteiro Paes, Yogesh Verma, Pin-Yu Chen, Steven Basart, Xinpeng Liu, Sebastian Gruber, Daoyuan Chen, Aurelie Lozano, Cenk Baykal, Milad Jalali Asadabadi, Muhammad Uzair Khattak, James M Murphy, Tianyu Cui, Pavlo Melnyk, Emanuele Natale, Eli Upfal, Minji Kim, Angio Kanazawa, Pavel Izmailov, Tao Song, Bang Liu, Munyque Mittelmann, Tim Roith, Mingtian Zhang, Oluwadamilola Fasina, Michael Felsberg, Sicong Huang, Steven Stenberg Hansen, Pan Xu, Biraj Dahal, Nicklas Hansen, Xikun Zhang, Francesco Belardinelli, Rui Chen, Jianguo Cao, Paulo Rauber, Yangjun Ruan, Xianqiang Lyu, Qiang Sheng, Spencer Peters, Alexey Luchinsky, Jiao Sun, Hanqing Zhu, Guoji Fu, Qingqing Zhao, Zhongyu Huang, Sihyun Yu, Suraj Kothawade, Luofeng Liao, Sergey Nikolenko, Ken Kahn, Qian Peng, Balaraman Ravindran, Panagiotis Karras, Ivan Serina, Giada Pistilli, Hyolim Kang, Hanchen Wang, Wenhao Wu, Peilin Zhou, Mehrtash Harandi, Jiaqi Han, Yin Fang, An Xu, Tom Bewley, Jiayi Wang, Franz Scherr, Zhengliang Liu, Yiqun Liu, Feng Liu, Difei Gao, Sixing Yu, Tsvetomila Mihaylova, Katherine M Collins, Siyi Tang, Zhipeng Liang, Samarth Mishra, Tyler H McCormick, Bohang Zhang, Bastian Rieck, Yangang Pan, Jack Zhang, Johannes Treutlein, Roman Levin, Xiangtai Li, Yangfeng Ji, Una-May O'Reilly, Lior Rokach, Yunjiang Jiang, Sung Ju Hwang, Eduardo Dadoalto, Bo Pan, Guendalina Righetti, Johannes Jakubik, Yanchao Sun, Hugo Berard, Eric Mitchell, Xuefeng Du, Seung-Bin Kim, Martin Burger, Yu Luo, Peter Macgregor, Nikolaus Howe, Erik Jenner, Florian J, Runlin Lei, Yan Zhang, Wei-Cheng Tseng, Craig Sherstan, Luojun Lin, Kai Yue, Anastasios Angelopoulos, Dzung Phan, Yuqi Bu, Junseok Lee, Safa Alver, Mengchen Zhang, Ali Seyfi, Chunyang Fu, Minghuan Liu, Enrico Scala, Neel Nanda, Shijie Liu, Roger B Grosse, Wen Song, Gautam Shroff, Mingming Gong, Yushun Dong, Yixin Zhang, Zijie Zhang, Ziqi Liu, Xingran Chen, Vincent Josua Hellendoom, Xiaoqing Tan, Kuan-Lin Chen, Dingfeng Shi, Simon Schierreich, Concetto Spampinato, Rob Brekelmans, Daniel D Lundstrom, Vahid Balazadeh Meresht, Viraj Prabhu, Zuxin Liu, Ronald J Brachman, Yung Chen, Liang Zeng, Georgios Birmapas, Xu Chen, Ayush Jain, Yuanwei Liu, Lawson

Wong, Muiyang Li, Junfeng Guo, Rayan Mazouz, Changho Shin, Yunjuan Wang, Charlie Dickens, Chongjian Ge, Zhijiang Guo, Oleg Vasilyev, Jean Barbier, Huifeng Yao, Luca Citi, Loay Mualem, Francesco Antici, Haozhi Qi, Flavio Miguel Varejao, Jiakuan Wang, Mohammad Azam Khan, Jun Bai, Vaden Masrani, Zhiwen Fan, Ian Stavness, Mingjian Wen, Qian Tao, Luyao Tang, Hanshu Yan, Chi Wang, Ulisses Braga-Neto, Hui-Po Wang, Anurag Arnab, Marc Law, Gabriele Corso, Pratyush Kumar, Shiyuan Chen, Zizheng Pan, Yujia Qin, Vaisakh Shaj, Jinfa Huang, Volker Gruhn, Marcus A Brubaker, Marco Favorito, Jiayuan Liu, Kartik Gupta, Zishun Yu, Aodong Li, Manolis Pitsikalis, Ivan F Rodriguez Rodriguez, Victoria Sherratt, Jaekyeom Kim, Felipe Meneguzzi, Chen-Lu Ding, Lionel Riou-Durand, Edward J Hu, Mingjie Sun, Jinheon Baek, Michael Witbrock, Mike Holenderski, Viraj Mehta, Freddie Kalaitzis, Vicenc Gomez, Yu Pan, Devansh Arpit, Deeksha Arya, Zipeng Bao, Prince Osei Aboagye, Tomasz Odrzygózd, Dimitar Iliev Dimitrov, Yasmeen Alsaedy, Alistair Muldal, Maksym Andriushchenko, Elena Zheleva, Rupesh Kumar Srivastava, Jinyoung Han, David Gros, Shangbin Feng, Hao-Tsung Yang, Nithin Nagaraj, Jung-Hun Kim, Leon Hetzel, Garrett Kenyon, Orr Paradise, Yuxuan Yi, Jason Xiaotian Dou, Andreas Loukas, Hyungjin Chung, Jin-Duk Park, Eric Zelikman, Sebastian Shenghong Tay, Jiajuan Tang, Kincaid Macdonald, Wonbin Kweon, Haoyue Dai, Achraf Azize, Calvin Tsay, Agnieszka Lawrynowicz, Qiyu Kang, Bingxin Zhou, Alena Shilova, Yuxiao Huang, Robert Honig, Sean Heigearthaig, Xovee Xu, Martin Gjoreski, Federico Berto, Shichao Xu, Zhigao Guo, Qian Pan, Wenjie Qiu, Charles X Ling, Krzysztof Maziarsz, Diane Oyen, Luigi Gresole, Yizhen Zheng, Changcheng Tang, Parth Patwa, Tomas Peitl, Anuroop Sriram, Dhruv Madeka, Hang Wang, John Aslanides, Hyeon Kyu Choi, Yongjun He, Joel Rixen, Krzysztof Sornat, Zhihan Gao, Renzhe Xu, Sebastian Stein, Lin Gu, Vianney Perchet, Javier De La Rosa, Luca Ragazzi, John M Dolan, Jinyu Cai, Krishnamurthy Kenthapadi, Alex Morehead, Tim Coleman, Animesh Garg, Jianhao Yan, Mihai Nica, Kaichun Mo, Maosen Li, Haoyu Chen, Dong-Hee Park, Johannes Oetsch, Yizhou Zhao, Seon-Ho Lee, Prasanna Parthasarathi, Haotian Bai, Shilong Liu, Zhiqi Huang, Jonathan Brophy, Vincent Stimper, Jinxin Liu, Kanji Uchino, William Harvey, Xu Luo, Ulrich Aivodji, Frederic Piedboeuf, Yang Tang, Joonhyun Jeong, Jeff Calder, Tao Wu, Wenjie Li, Stefano Panzeri, Kunal Pratap Singh, Cynthia Matuszek, Ke Sun, Ching-Yun Ko, Micah Goldblum, Jiayu Chen, Geon Lee, Zhiyuan Fang, Chi-Ming Chung, Zixia Jia, Peng Ye, Jędrzej Potoniec, Yash Akhauri, Chung-Yeon Lee, Asif Khan, Thomas Meyer, Pablo Barcelo, Salman Khan, Anna Ivanova, Dzmitry Tishchou, Gabriel Kreiman, Fang-Lue Zhang, Jingkai Zhu, Minchul Shin, Ross M Clarke, Qizhang Feng, Fuqi Jia, Nicolas Troquard, Desik Rengarajan, Sebastian Dalleiger, Selena Zihan Ling, Phil Pope, Beier Zhu, Murat Kantarcioglu, Dianwen Ng, Chunwei Ma, Yi Wu, Qihang Lin, Shaunak Srivastava, Zhuo Li, Muxuan Liang, Tsuyoshi Ide, Kenny Young, Umang Bhatt, Nicolas Lanzetti, Nayeon Lee, Mengchu Li, Joan Puigcerver, Lingjie Mei, Woosook Shin, Xiaokang Chen, Junpei Komiyama, Dhruv Batra, Melanie Sclar, Marta Kryven, Yihan Zhang, Matthias Weissenbacher, Diego Granziol, Saurabh Sihag, Batu Ozturkler, Eduard Hovy, Yuning You, Ruizhong Qiu, Jiaru Zhang, Anshuman Chhabra, Didier Chetelat, Joey Bose, Raphael Poulain, Ben Armstrong, Xin Wen, Youngmoon Lee, Daniel Karl I Weidele, Tom Dhaene, Tom Drummond, Sangwoo Mo, Anh Tong, Haojiang Li, Xu Cao, Gergely Flamich, Yonggang Zhang, Yao-Yuan Yang, Hiroki Furuta, Shitao Tang, Huang Bojun, Jiaming Zhu, Hongyao Tang, Benjamin Guedj, Cade Gordon, Tao He, Daria Stepanova, Li Chen, Timothy Hospedales, Gaetan Marceau Caron, Mancheng Meng, Fabien Gandon, Son Tran, Paternie Gahungu, Elliot Paquette, Lauro Langosco Di Langosco, Zhuoyue Lyu, Zhixiang Wei, Gaurav Pandey, Ziheng Duan, Elisa Fromont, Stefan Heidekruger, Borja G Leon, Meng Liu, Mira Welner, Gang Yan, Giovanni Chierchia, Rasmus Pagh, Thomas Pierrot, Patryk Wielopolski, Ivan Marisca, Matthias Samwald, Yangfu Zhu, Zuo Chen, Yining Ma, Joseph Early, Luziwei Leng, Laura Wynter, Louis Bethune, Vincent Mai, Matthew I Swindall, Julia Grabinski, John Burden, Ankit Vani, Guillaume Leclerc, Tianying Ji, Soroush Vosoughi, Fermin Travi, Hai Zhao, Xuan Di, Ruofan Liang, Gregor Bachmann, Pradeep Ravikummar, Taoan Huang, Weina Jin, Jiayuan Mao, Deepak Ramachandran, Xiangzhe Kong, Ziyao Huang, Yuko Ishiwaka, Giancarlo Kerg, Joanna Hong, Ziheng Zeng, Seong Joon Oh, Renrui Zhang, Abhishek Gupta, Cole M Wyeth, Jinguo Zhu, Peter Kairouz, Jian Zhao, Soren Mindermann, Bastien Maubert, Longhui Yu, Sujun Dutta, Balaji Vasan Srinivasan, Michael Littman, Mariusz Oszust, Edouard Yvinec, Nishanth Dikkala, Kate Sanders, Hongzhi Wu, Emir Ceyani, Geraud Nangue Tasse, Georgios Georgakis, Zhenfeng He, Hyunwoo J Kim, Danyal Saeed, Bhumika Mishra, Sirisha Rambhatla, Canzhe Zhao, Zulong Chen, Louis Tiao, Marco Console, Tahar Allouche, Seungyong Moon, Daniel Mackinlay, Jing Liu, Kuan Yew Leong, Qingbiao Li, Anwen Hu, Arman Zharmagambetov, Jiawei Shao, Andrew Li, Luke E Richards, Aurelien Garivier, Danda Pani Paudel, Riccardo Marin, Chih-Kuan Yeh, Chao Qin, Ye Zhu, Yinglin Duan, I-Hsiang Chen, Kartik Chandra, Yi Zeng, John E Laird, Leonardo Amado, Luis Fernando D'Haro, Zeyu Qin, Santu Rana, Jiajuan Fan, Joshua Welch, Adam Prugel-Bennett, Tristan Bepler, Tai Sing Lee, Georg Ostrovski, Mohammed Haroon Dupty, Niclas Boehmer, Yinkai Wang, Fabian Paischer, Sihan Liu, Drew Linsley, Wuwei Lin, Ziyi Chen, Carol Long, Seongku Kang, Haoyuan Chen, Danyang Tu, Vaibhav V Unhelkar, Ivan Orsolich, Ivan Donadello, Jacob Euler-Rolle, Mohan Zhang, Matej Racinsky, Jiaqing Liang, Ravinder Bhattoo, Javier Segovia-Aguas, Yifan Yang, Isabelle Augenstein, David Lindner, Keane Lucas, Huan Zhang, Hyung Ju Suh, Nikolaos Tsilivis, Pandeng Li, Evgeny S Saveliev, Wendi Li, Shengchao Liu, Pierre Tassel, Avi Mohan, Jinjin Gu, Lapo Faggi, Zhen Zhu, Vihang Patil, Giuseppe Vietri, Morten Goodwin, Zizheng Yan, Erik Daxberger, Yaqian Zhang, Michael Niemeyer, Mauricio Gruppi, Yipeng Qin, Amir Rahmati, Pietro Mazzaglia, Yiwen Kou, Karan Singhal, Fatih Erdem Kizilkaya, Xiangyi Chen, Haoxuan Qu, Lei Jiao, Emanuele Marconato, Zhen-Duo Chen, Cecilia G Morales, James P Delgrande, Zhenyuan Zhou, Vasu Singla, Qian Chen, Guojun Xiong, Peter Stone, Blake Richards, James M Murray, Yu Huang, Tim Verbelen, Edward De Brouwer, Nur Muhammad Shafullah, Zhenke Wu, Konstantinos Voumouris, Nuo Xu, Youngkyoung Kim, Teemu Roos, Mohammad Reza Belbasi, Wenda Li, Shengyang Sun, Olivier Graffeuille, Marc Dymetman, Yunzhe Li, Markus Hiller, Zhenhailong Wang, Filip Cano Cordoba, Xingjian Du, Zongzhang Zhang, Sandeep Silwal, Mengping Yang, Fabricio Vaselali, Aaron Defazio, Samuel Yang-Zhao, Roberto Posenato, Alexis Joly, Steven Yin, Pedro Bizarro, Chujie Zheng, Subrata Mitra, Neel Garg, Jun Fang, Emanuele Guidotti, Roy Fox, Weiming Liu, Steve Peterson, Harsh M Patel, Dan Navon, Romain Mueller, Rupert Freeman, Ilya Chugunov, Spyridon Samothrakis, Andrea Cini, Kumar Kshitij Patel, Shadi Albarqouni, Xian Yeow Lee, Yichong Leng, Junkyu Lee, Hezhen Hu, Antoine Bosselut, Udayan Khurana, Wee Sun Lee, Dino Sejdinovic, Jiancheng Yang, Ana Bazzan, Mario A T Figueiredo, Eunggu Yun, Marco Roveri, Frederik Schubert, Liushuai Shi, Misoo Kim, Ziyang Song, Xiaoyang Wu, Fabrizio Silvestri, Alan Yuille, Xizewen Han, Mykel J Kochenderfer, Keshav Bhandari, Fan Lyu, Sehyun Hwang, Sheng Li, Yuzheng Hu, Todd W Neller, Wanhua Li, Philipp Tholke, Jason Hartford, Sunwoo Lee, Weiqiang Zheng, Yi Cheng, Jithin Jagannath, Yangrui Chen, Jie Yang, Anders Karlsson, Hongsheng Hu, Zhenbin Wang, Xiaohao Xu, Carlos Escolano, Khai Nguyen, Zhen Dong, Helge Spieker, Honghao Wei, Mingzhi Dong, Zhiyu Lin, Ilsang Ohn, Jiayi Ying, Namgyu Ho, Stephan Hoyer, Juan Esteban Kamienkowski, Haoran Liu, Yinan Feng, Lingfeng Sun, Mingyuan Zhang, Songhua Wu, Zhichao Wang, Stefano Leonardi, Jose Hernandez-Orallo, Chenyang Wu, He Sun, Adrian Feiguin, Chaoyue Liu, Xiaohua Wu, Yiren Zhao, Nicholas Roberts, Zhaowei Zhu, Fei Huang, Gemine Vivone, Scott Fujimoto, Kerstin Bach, Samuel I Holt, Ashok Narendranath, Elliot Nelson, Zhenyu Yang, Safa C Medin, Mark Crowley, Liyuan Wang, Maria Hedblom, Alfio Ferrara, Pratik Vaishnavi, Ernesto Evgeniy Sanchez Shayda, Mireille El Gheche, Ariel Kwiatkowski, Oliver Kiss, Bingchen Zhao, Eddie Ubri, Mert Yuksekgonul, Minyang Hu, Jessica Shi, Changan Niu, Yoav Wald, Maxwell Xu, Nadjib Lazaar, Liangli Zhen, Yaochen Xie, Erdem Biyik, Tian Zhou, Ye Yuan.

8 Funding

This research project was funded by Open Philanthropy, 182 Howard Street #225. San Francisco, CA 94105, USA.

9 Additional Information

Correspondence and requests for materials should be addressed to Katja Grace.

Reprints and permissions requests should be directed to Katja Grace.

References

- Katja Grace, Zach Stein-Perlman, Ben Weinstein-Raun, and John Salvatier. 2022 Expert Survey on Progress in AI. *AI Impacts*, Aug 2022. URL https://wiki.aiimpacts.org/ai_timelines/predictions_of_human-level_ai_timelines/ai_timeline_surveys/2022_expert_survey_on_progress_in_ai.
- OpenAI. Moving AI governance forward, Jul 2023. URL <https://openai.com/blog/moving-ai-governance-forward>.

- Center for Human-compatible Artificial Intelligence. Research Publications – Center for Human-Compatible Artificial Intelligence, 2023. URL <https://humancompatible.ai/research>.
- Gavin Newsom. EXECUTIVE ORDER N-12-23, Sep 2023. URL <https://www.gov.ca.gov/wp-content/uploads/2023/09/AI-E0-No.12--GGN-Signed.pdf>.
- AI.gov. Making AI Work for the American People, 2023. URL <https://ai.gov>.
- Inter-Agency Working Group on Artificial Intelligence. Principles for the Ethical Use of Artificial Intelligence in the United Nations System (Advanced unedited version), September 2022.
- Richard Korzekwa and Harlan Stewart. Views of prominent ai developers on risk from ai, 2023. URL https://wiki.aiimpacts.org/arguments_for_ai_risk/views_of_ai_developers_on_risk_from_ai.
- Pablo Villalobos. Scaling laws literature review, 2023. URL <https://epochai.org/blog/scaling-laws-literature-review>. Accessed: 2023-12-22.
- Katja Grace, Rick Korzekwa, Asya Bergal, and Daniel Kokotajlo. Discontinuous Progress Investigation. Technical report, AI Impacts, 2021. URL https://wiki.aiimpacts.org/ai_timelines/discontinuous_progress_investigation.
- Stephen M. Omohundro. The basic ai drives. In *Proceedings of the 2008 Conference on Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, page 483–492, NLD, 2008. IOS Press. ISBN 9781586038335.
- Peter S Park, Simon Goldstein, Aidan O’Gara, Michael Chen, and Dan Hendrycks. Ai deception: A survey of examples, risks, and potential solutions. *arXiv preprint arXiv:2308.14752*, 2023.
- Charles I Jones. The ai dilemma: Growth versus existential risk. Technical report, National Bureau of Economic Research, 2023.
- Philip Trammell and Anton Korinek. Economic growth under transformative ai. Technical report, National Bureau of Economic Research, 2023.
- Future of Life Institute. Pause Giant AI Experiments: An Open Letter, 2023. URL <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>.
- Center for AI Safety. Statement on AI Risk: AI experts and public figures express their concerns about AI risk., 2023. URL <https://www.safe.ai/statement-on-ai-risk>.
- Joseph R Biden. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 2023. URL <https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/>.
- Dario Amodei. Written testimony of dario amodei, ph.d. co-founder and ceo, anthropic; for a hearing on “oversight of a.i.: Principles for regulation”; before the judiciary committee subcommittee on privacy, technology, and the law; united states senate, Jul 2023. URL https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-testimony_-_amodei.pdf.
- gov.uk. AI Safety Summit 2023 - GOV.UK, November 2023. URL <https://www.gov.uk/government/topical-events/ai-safety-summit-2023>.
- European Parliament. EU AI Act: first regulation on artificial intelligence. *Accessed June, 25: 2023*, 2023. URL <https://www.europarl.europa.eu/news/en/headlines/society/20230601ST093804/eu-ai-act-first-regulation-on-artificial-intelligence>.
- Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. When will ai exceed human performance? evidence from ai experts. *Journal of Artificial Intelligence Research*, 62:729–754, 2018a.
- Zachary Stein-Perlman. Surveys of US public opinion on AI, 2023. URL https://wiki.aiimpacts.org/responses_to_ai/public_opinion_on_ai/surveys_of_public_opinion_on_ai/surveys_of_us_public_opinion_on_ai.
- Michael Chui, Lareina Yee, Bryce Hall, and Alex Singla. The state of AI in 2023: Generative AI’s breakout year. Technical report, McKinsey & Company, 2023.
- AI Impacts. 2023 Expert Survey on Progress in AI [Survey PDF], 2023a. URL https://wiki.aiimpacts.org/_media/ai_timelines/predictions_of_human-level_ai_timelines/ai_timeline_surveys/2023_espai_paid.pdf.
- Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *science*, 211(4481): 453–458, 1981.

- AI Impacts. 2023 Expert Survey on Progress in AI [AI Impacts Wiki], 2023b. URL https://wiki.aiimpacts.org/ai_timelines/predictions_of_human-level_ai_timelines/ai_timeline_surveys/2023_expert_survey_on_progress_in_ai.
- O*NET. All Job Family Occupations, 2023. URL <https://www.onetonline.org/find/family?f=0>.
- Joseph Carlsmith. Is Power-Seeking AI an Existential Risk? *arXiv preprint arXiv:2206.13353*, 2022.
- Stuart Russell. Of myths and moonshine, 2014. URL <https://www.edge.org/conversation/the-myth-of-ai#26015>.
- Philip E. Tetlock. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press, Princeton, 2005. ISBN 9781400888818. doi:10.1515/9781400888818. URL <https://doi.org/10.1515/9781400888818>.
- Tamara Savage, Alex Davis, Baruch Fischhoff, and M Granger Morgan. A strategy to improve expert technology forecasts. *Proceedings of the National Academy of Sciences*, 118(21):e2021558118, 2021.
- James Surowiecki. *The Wisdom of Crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations*. Doubleday & Co., 2004.
- Ezra Karger, Josh Rosenberg, Zachary Jacobs, Molly Hickman, Rose Hadshar, Kayla Gamin, Taylor Smith, Tegan McCaslin, Stephen Thomas, and Philip E Tetlock. Forecasting existential risks: Evidence from a long-run forecasting tournament, 2023.
- Michael Braun Hamilton. Online survey response rates and times: Background and guidance for industry. *Tercent, Inc*, 2003.
- National Research Council (US) Committee on Assessing Fundamental Attitudes of Life Scientists as a Basis for Biosecurity Education. A survey of attitudes and actions on dual use research in the life sciences: A collaborative effort of the national research council and the american association for the advancement of science. Technical report, American Association for the Advancement of Science and National Research Council and others, 2009.
- Fabienne Sandkühler, Harlan Stewart, Richard C Korzekwa, and Jan Brauner. 2023 Expert Survey on Progress in AI, Oct 2023. URL <https://osf.io/8gzdr>.
- AI Impacts. AI Timeline Surveys, 2023c. URL https://wiki.aiimpacts.org/ai_timelines/predictions_of_human-level_ai_timelines/ai_timeline_surveys/ai_timeline_surveys.
- Philippe Aghion, Benjamin F Jones, and Charles I Jones. Artificial intelligence and economic growth. Technical report, National Bureau of Economic Research, 2017.
- Dennis Bray and Hans von Storch. Climate science: An empirical example of postnormal science. *Bulletin of the American Meteorological Society*, 80(3):439–456, 1999.
- Dennis Bray and Hans Von Storch. *CliSci2008: A survey of the perspectives of climate scientists concerning climate science and climate change*. GKSS-Forschungszentrum Geesthacht Geesthacht, 2010.
- Peter T Doran and Maggie Kendall Zimmerman. Examining the scientific consensus on climate change. *Eos, Transactions American Geophysical Union*, 90(3):22–23, 2009.
- Katja Grace, John Salvatier, Allen Dafoe, Baobao Zhang, and Owain Evans. Viewpoint: when will AI exceed human performance? Evidence from AI Experts. *Journal of Artificial Intelligence Research*, pages 729–754, 2018b. doi:10.1613/jair.1.11222.
- Allyson L Holbrook, Jon A Krosnick, and Alison Pfent. The causes and consequences of response rates in surveys by the news media and government contractor survey research firms. *Advances in telephone survey methodology*, pages 499–528, 2007.
- Thomas R Stewart. Scientists’ uncertainty and disagreement about global climate change: A psychological perspective. *International Journal of Psychology*, 26(5):565–573, 1991.
- Amos Tversky and Daniel Kahneman. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, 90(4):293, 1983.
- Elena Verdolini, Laura Diaz Anadon, Erin Baker, Valentina Bosetti, and Lara Reis. The future prospects of energy technologies: Insights from expert elicitations. *Review of Environmental Economics and Policy*, 12(1), 2016.
- Penny S Visser, Jon A Krosnick, Jesse Marquette, and Michael Curtin. Mail surveys for election forecasting? an evaluation of the columbus dispatch poll. *Public Opinion Quarterly*, 60(2):181–227, 1996.
- Ryan Wiser, Joseph Rand, Joachim Seel, Philipp Beiter, Erin Baker, Eric Lantz, and Patrick Gilman. Expert elicitation survey predicts 37% to 49% declines in wind energy costs by 2050. *Nature Energy*, 6(5):555–565, 2021.

Baobao Zhang, Noemi Dreksler, Markus Anderljung, Lauren Kahn, Charlie Giattino, Allan Dafoe, and Michael C Horowitz. Forecasting ai progress: Evidence from a survey of machine learning researchers. *arXiv preprint arXiv:2206.04132*, 2022.

A Results Comparing Different Demographics

A.1 How soon will High-level Machine Intelligence be feasible?

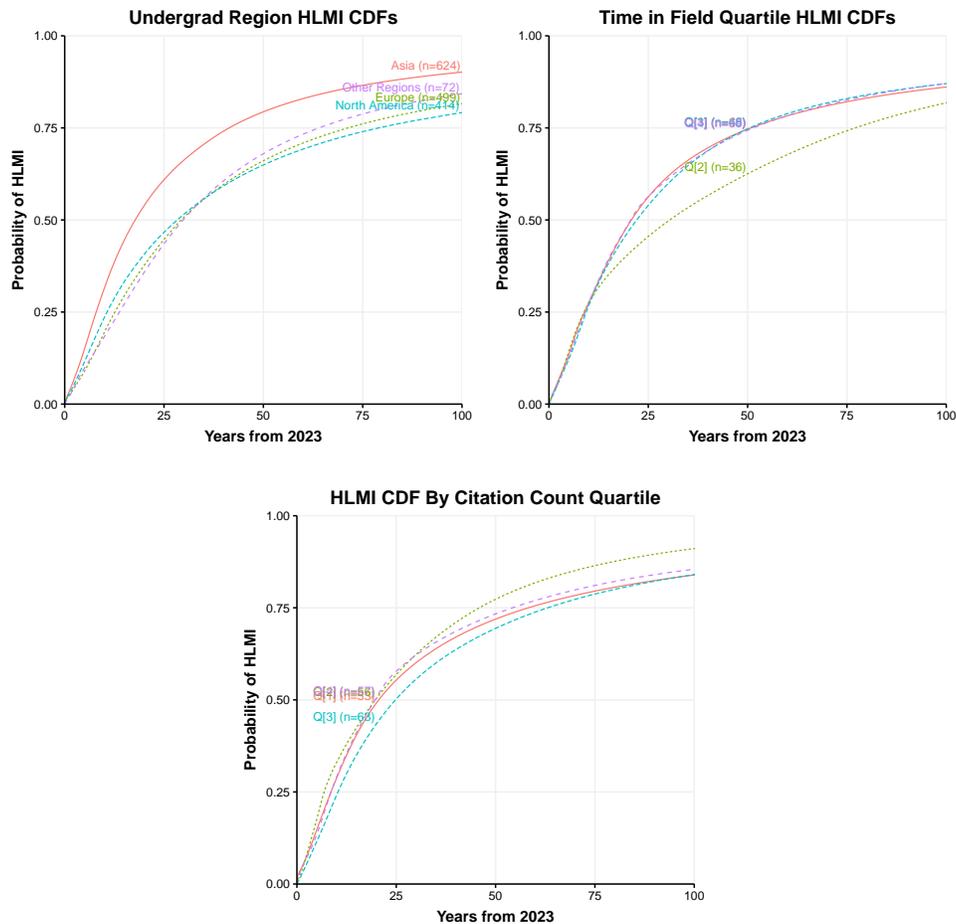


Figure 17: Aggregate forecasts for time until HLMI were shorter for participants whose region of undergraduate study was Asia. Time in field did not have a significant effect on forecasts for time until HLMI. Citation count did not have a significant effect on forecasts for time until HLMI.

A.2 How good or bad for humans will HLMI be?

A.2.1 Amount of thinking about the issue

We asked all respondents how much thought they have given to the social impacts of smarter-than-human machines.

It is not clear how to interpret the results of this question. Specifically, while thinking more about a topic presumably improves predictions about it, people who think a lot may do so because they are concerned, so the association could also be due to this selection effect.

	“Very little” or “a little” thought (n = 621)	“A lot” or “a great deal” of thought (n = 1113)
Percentage of these respondents who thought that the probability of an extremely good outcome was greater than 10%	68%	41%
Percentage of these respondents who thought that the probability of an extremely bad outcome was greater than 10%	34%	71%

Table 4: **People who had thought more about the social impacts of smarter-than-human machines were substantially less likely to give more than 10% credence to extremely optimistic outcomes, and substantially more likely to give extremely negative outcomes more than 10% credence.**

A.2.2 Academia vs industry, undergraduate region, and amount of thought

These are the demographic differences within the answers to the question about how good or bad HLMI will be overall.

	Extremely good	On balance good	Neutral	On balance bad	Extremely bad
Industry	12.75	30	20	15	5
Academia	10	25	20	15	5
Asia	15	30	20	10	5
North America	10	25	20	15	5
Other location	10	25	20	17.5	5
Having thought more	15	25	20	15	5
Having thought less	10	30	20	12	5

Table 5: **Median probabilities given to each outcome by each demographic group.**

A.3 How likely is AI to cause human extinction?

Whether participants worked in academia or industry did not affect median responses much—both had a median of 5%. The region respondents graduated from affected responses somewhat: the Asian median was 10%, while North American and European medians were 5%. Having thought more (either “a lot” or “a great deal”) about the question was associated with a median of 9%, while having thought “little” or “very little” was associated with a median of 5%.

The similarity of answers across several slightly different questions, across this and previous surveys, and across participants in academia and industry and different geographic regions appears to be robust evidence that the majority of AI researchers think there is a nontrivial risk of extinction or similar catastrophes due to AI.

B Supplementary Figures

B.1 How soon will human-level performance on all tasks or occupations be feasible?

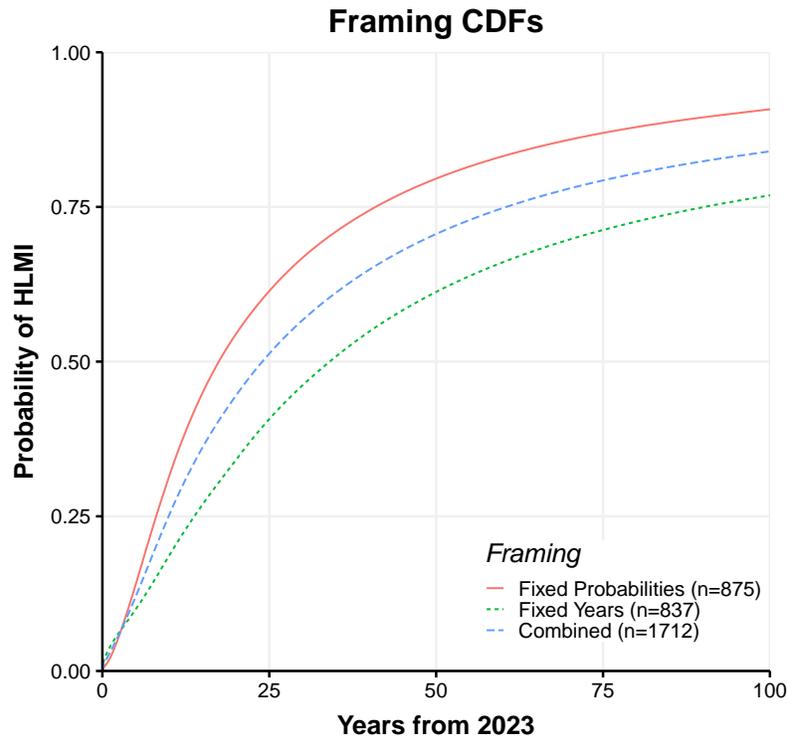


Figure 18: Participants who received questions framed in terms of fixed-years had later forecasts than those who received questions in terms of fixed-probabilities.

B.2 How soon will Full Automation of Labor be feasible?

	2016 aggregate forecast	2022 aggregate forecast	2023 aggregate forecast
Year with a 50% chance of HLMI	2061 (n*=259)	2060 (n=461)	2047 (n=1714)
Year with a 10% chance of HLMI	2025 (n*=259)	2029 (n=461)	2027 (n=1714)
Year with a 50% chance of FAOL	2138 (n*=97)	2164 (n=202)	2116 (n=774)
Year with a 10% chance of FAOL	2036 (n*=97)	2050 (n=202)	2037 (n=774)

Table 6: **By what year will human-level performance on all tasks (HLMI) or occupations (FAOL) be feasible?** Aggregate forecasts for HLMI and FAOL in 2023 have gotten earlier since 2022. Comparing 2023 to 2016, 2023's 50% estimates are earlier, but the 10% estimates are later. *n reported for 2016 only is total responses rather than valid responses after cleaning.

B.3 Will there be an intelligence explosion?

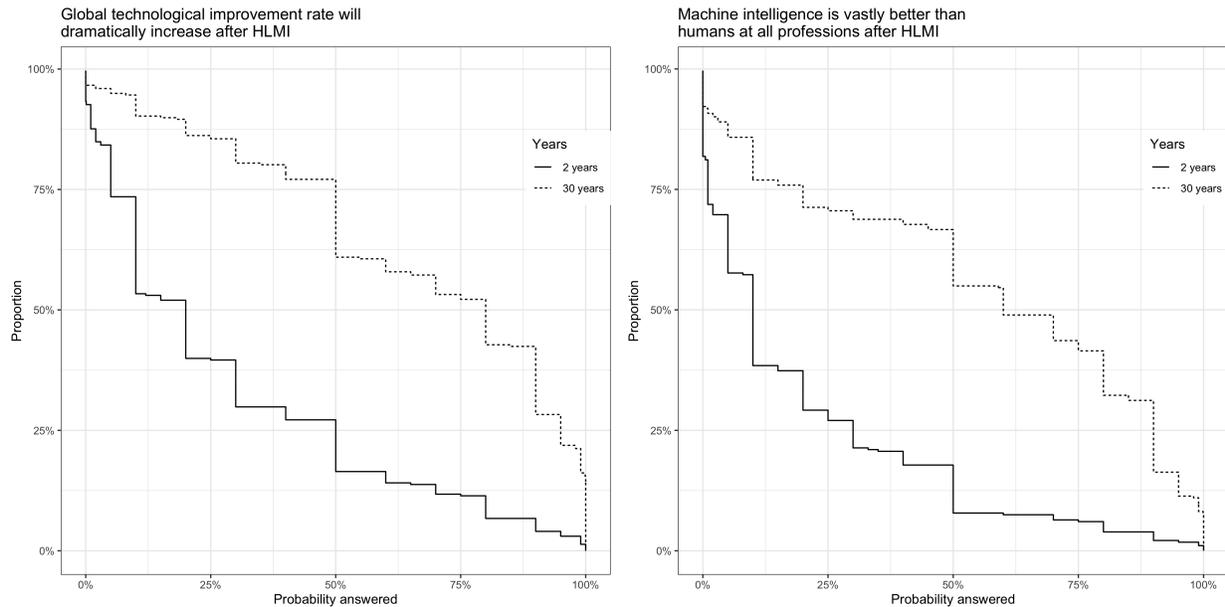


Figure 19: **Two framings of the intelligence explosion question:** How likely is an “explosive global technological improvement” two and 30 years after HLMI? And, How likely is it that AI be “vastly better” than humans in all tasks two and 30 years after HLMI? In the first framing, the median prediction for 2 years post-HLMI was 20%, whereas the median prediction for 30 years post-HLMI was 80%. In the second framing, the median prediction for 2 years post-HLMI was 10%, whereas the median prediction for 30 years post-HLMI was 60%.

B.4 How much should AI safety research be prioritized?

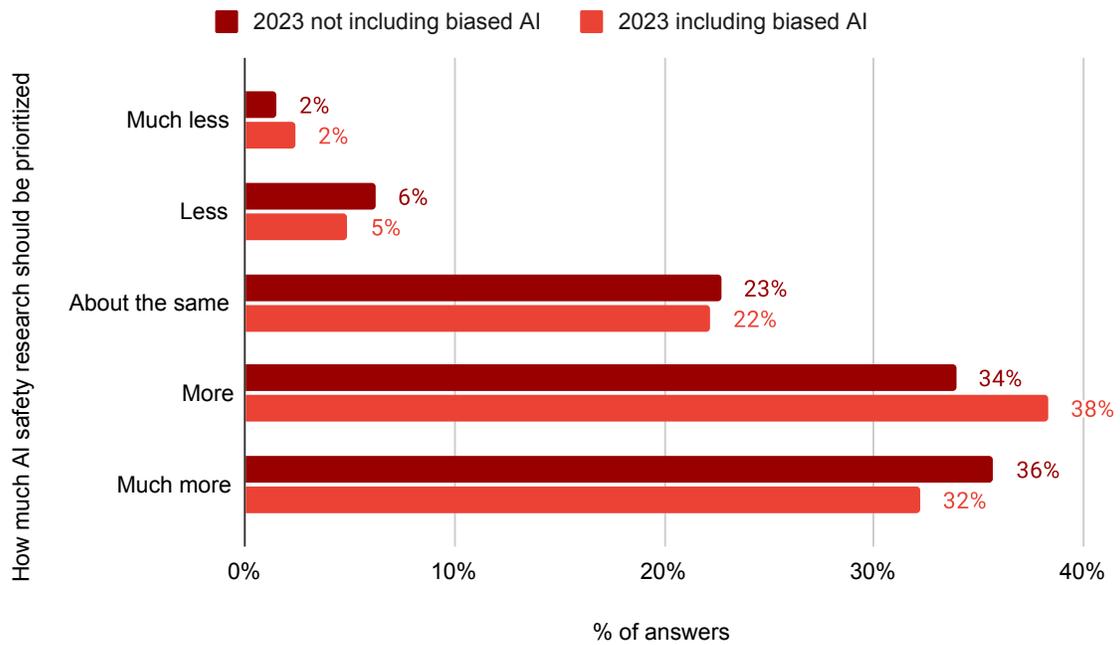


Figure 20: Two framings of the question “How much should AI safety research be prioritized?”, one including and one not including biased AI as an example.

C Full Descriptions of Milestones

This list is also available at <https://tinyurl.com/aitasks>.

Short name	Full description from question
Physically install wiring in a house	Given a one-sentence description of the task and given the same information you would give a human to perform this task (such as information about the house), physically install the electrical wiring in a new home, without more input from humans.
Finetune LLM	Given a one-sentence description of the task, download and finetune an existing open source LLM, without more input from humans. The fine-tune must improve the performance of the LLM on some predetermined benchmark metric.
Find and patch security flaw	Given a one-sentence description of the task and no more input from humans, find and patch a security flaw in an open source project with over 100,000 users.
Run ML study and write paper	Given a one sentence description of a research question in machine learning, conduct a study that would inform the answer to that question and write a paper of a quality that could be accepted at a leading machine learning conference, without more input from humans.
Replicate ML paper	Given a study published at a leading machine learning conference, replicate the study without more input from humans. The replication must meet the standards of the ML Reproducibility Challenge (https://paperswithcode.com/rc2022).
Build payment processing website	Given a set of specifications, build a website from scratch that can handle payment processing, including the frontend, backend, and secure payment integration, without more input from humans.
Long unsolved math problem	Given a list of long-standing unsolved problems in mathematics, such as the Millennium Prize problems or one the problems in “Unsolved Problems on Mathematics for the 21st Century,” solve one without more input from humans.
Translate text from a newly discovered language using a bilingual document	Translate a text written in a newly discovered language into English as well as a team of human experts, using a single other document in both languages (like a Rosetta stone). Suppose all of the words in the text can be found in the translated document, and that the language is a difficult one.
Translate new language speech using films with subtitles and existing bilingual data	Translate speech in a new language given only unlimited films with subtitles in the new language. Suppose the system has access to training data for other languages, of the kind used now (e.g. same text in two languages for many languages and films with subtitles in many languages).
Translate text nearly as well as a bilingual but untrained translator	Perform translation about as good as a human who is fluent in both languages but unskilled at translation, for most types of text, and for most popular languages (including languages that are known to be difficult, like Czech, Chinese and Arabic).
Offer phone banking services, including unique tasks, on par with human operators	Provide phone banking services as well as human operators can, without annoying customers more than humans. This includes many one-off tasks, such as helping to order a replacement bank card or clarifying how to use part of the bank website to a customer.
Classify unseen objects into categories after training on different but similar classes	Correctly group images of previously unseen objects into classes, after training on a similar labeled dataset containing completely different classes. The classes should be similar to the ImageNet classes.

Recognize a new object in varied settings after seeing it just once	<p>One-shot learning: see only one labeled image of a new object, and then be able to recognize the object in real world scenes, to the extent that a typical human can (i.e. including in a wide variety of settings). For example, see only one image of a platypus, and then be able to recognize platypuses in nature photos. The system may train on labeled images of other objects.</p> <p>Currently, deep networks often need hundreds of examples in classification tasks¹, but there has been work on one-shot learning for both classification² and generative tasks.³</p> <p>1: Lake et al. (2015). Building Machines That Learn and Think Like People 2: Koch (2015). Siamese Neural Networks for One-Shot Image Recognition 3: Rezende et al. (2016). One-Shot Generalization in Deep Generative Models</p>
Create a 3D model and realistic video from a new angle of a scene	<p>See a short video of a scene, and then be able to construct a 3D model of the scene good enough to create a realistic video of the same scene from a substantially different angle. For example, constructing a short video of walking through a house from a video taking a very different path through the house.</p>
Transcribe speech in noise and varied accents on par with humans	<p>Transcribe human speech with a variety of accents in a noisy environment as well as a typical human can.</p>
Output written text as a recording indistinguishable from a voice actor	<p>Take a written passage and output a recording that can't be distinguished from a voice actor, by an expert listener.</p>
Prove and generate math theorems publishable in top journals	<p>Routinely and autonomously prove mathematical theorems that are publishable in top mathematics journals today, including generating the theorems to prove.</p>
Win Putnam Math Competition (problems with known but very difficult answers)	<p>Perform as well as the best human entrants in the Putnam competition—a math contest whose questions have known solutions, but which are difficult for the best young mathematicians.</p>
Beat humans at Go (after same # games of training)	<p>Defeat the best Go players, training only on as many games as the best Go players have played.</p> <p>For reference, DeepMind's AlphaGo has probably played a hundred million games of self-play, while Lee Sedol has probably played 50,000 games in his life.¹</p> <p>1: Lake et al. (2015). Building Machines That Learn and Think Like People</p>
Beat ($\geq 50\%$ games) best Starcraft 2 players given only video of screen	<p>Beat the best human Starcraft 2 players at least 50% of the time, given a video of the screen.</p> <p>Starcraft 2 is a real time strategy game characterized by:</p> <ul style="list-style-type: none"> • Continuous time play • Huge action space • Partial observability of enemies • Long term strategic play, e.g. preparing for and then hiding surprise attacks.
Match human novice in any new computer game in <10 min	<p>Play a randomly selected computer game, including difficult ones, about as well as a human novice, after playing the game less than 10 minutes of game time. The system may train on other games.</p>
Outperform humans in new Angry Birds levels	<p>Play new levels of Angry Birds better than the best human players. Angry Birds is a game where players try to efficiently destroy 2D block towers with a catapult. For context, this is the goal of the IJCAI Angry Birds AI competition.¹</p> <p>1: aibirds.org</p>
Beat pro testers in all Atari games	<p>Outperform professional game testers on all Atari games using no game-specific knowledge. This includes games like Frostbite, which require planning to achieve sub-goals and initially posed problems for deep Q-networks.^{1,2}</p> <p>1: Mnih et al. (2015). Human-level control through deep reinforcement learning 2: Lake et al. (2015). Building Machines That Learn and Think Like People</p>

Beat novices in 50% of Atari games after 20 min play	<p>Outperform human novices on 50% of Atari games after only 20 minutes of training play time and no game specific knowledge.</p> <p>For context, the original Atari playing deep Q-network outperforms professional game testers on 47% of games,¹ but used hundreds of hours of play to train.²</p> <p>1: Mnih et al. (2015). Human-level control through deep reinforcement learning 2: Lake et al. (2015). Building Machines That Learn and Think Like People</p>
Fold laundry as well and as fast as the median human clothing store employee	Fold laundry as well and as fast as the median human clothing store employee.
Beat fastest human runners in a 5km city streets race using bipedal robot body	Beat the fastest human runners in a 5 kilometer race through city streets using a bipedal robot body.
Build any LEGO set using using non-specialized robotics, with instructions	<p>Physically assemble any LEGO set given the pieces and instructions, using non-specialized robotics hardware.</p> <p>For context, Fu 2016¹ successfully joins single large LEGO pieces using model based reinforcement learning and online adaptation.</p> <p>1: Fu et al. (2016). One-Shot Learning of Manipulation Skills with Online Dynamics Adaptation and Neural Network Priors</p>
Efficiently sort large number lists beyond training size	<p>Learn to efficiently sort lists of numbers much larger than in any training set used, the way Neural GPUs can do for addition,¹ but without being given the form of the solution.</p> <p>For context, the original Neural Turing Machines could not do this,² but Neural Programmer-Interpreters³ have been able to do this by training on stack traces (which contain a lot of information about the form of the solution).</p> <p>1: Kaiser & Sutskever (2015). Neural GPUs Learn Algorithms 2: Zaremba & Sutskever (2015). Reinforcement Learning Neural Turing Machines 3: Reed & de Freitas (2015). Neural Programmer-Interpreters</p>
Write readable Python code for algorithms like quicksort from specs and examples	<p>Write concise, efficient, human-readable Python code to implement simple algorithms like quicksort. That is, the system should write code that sorts a list, rather than just being able to sort lists.</p> <p>Suppose the system is given only:</p> <ul style="list-style-type: none"> • A specification of what counts as a sorted list • Several examples of lists undergoing sorting by quicksort
Answer Googleable factoid questions better than expert (w/ web)	<p>Answer any “easily Googleable” factoid questions posed in natural language better than an expert on the relevant topic (with internet access), having found the answers on the internet.</p> <p>Examples of factoid questions:</p> <ul style="list-style-type: none"> • “What is the poisonous substance in Oleander plants?” • “How many species of lizard can be found in Great Britain?”
Answer Googleable but open-ended factual questions better than expert (w/ web)	<p>Answer any “easily Googleable” factual but open ended question posed in natural language better than an expert on the relevant topic (with internet access), having found the answers on the internet.</p> <p>Examples of open ended questions:</p> <ul style="list-style-type: none"> • “What does it mean if my lights dim when I turn on the microwave?” • “When does home insurance cover roof replacement?”

Answer undecided factual questions	Give good answers in natural language to factual questions posed in natural language for which there are no definite correct answers. For example: “What causes the demographic transition?”, “Is the thylacine extinct?”, “How safe is seeing a chiropractor?”
Write high-grade, unique high school history essays without plagiarizing	Write an essay for a high-school history class that would receive high grades and pass plagiarism detectors. For example answer a question like “How did the whaling industry affect the industrial revolution?”
Create songs that can hit the US Top 40 (full audio file)	Compose a song that is good enough to reach the US Top 40. The system should output the complete song as an audio file.
Fake new song indistinguishable from a specific artist’s work by expert listeners	Produce a song that is indistinguishable from a new song by a particular artist, e.g. a song that experienced listeners can’t distinguish from a new song by Taylor Swift.
Write a novel or story that could land on the NYT best-seller list	Write a novel or short story good enough to make it to the New York Times best-seller list.
Concisely and completely explain AI’s computer game moves	For any computer game that can be played well by a machine, explain the machine’s choice of moves in a way that feels concise and complete to a layman.
Play poker well enough to win the World Series of Poker	Play poker well enough to win the World Series of Poker.
Deduce and symbolize physical laws (e.g. Newtonian mechanics) of a virtual world	After spending time in a virtual world, output the differential equations governing that world in symbolic form. For example, the agent is placed in a game engine where Newtonian mechanics holds exactly and the agent is then able to conduct experiments with a ball and output Newton’s laws of motion.

D Additional Information on Participation Bias and Question-level Response Bias

As with any survey, our results could be skewed by participation bias, if participants had systematically different opinions than those who chose not to participate. Here we review evidence about the presence of participation bias. We find no evidence suggesting strong participation bias.

We sought to minimize participation bias in several ways. First, we made efforts to increase the response rate in ways we expected to attract broad participation not particularly correlated with opinion. For instance, we paid participants \$50 or an equivalent reward for taking the survey (which we estimated to take a median 16 minutes). We also sent a pre-notification email, in it invited questions, and discussed concerns with researchers who wrote. We also sent at least four reminders to take the survey.

As well as aiming to attract a large and broad set of respondents, we tried to limit the ability of recipients to choose to enter the survey based on their opinions, by limiting cues about the survey content before entering the survey. In particular we said the topic was ‘the future of AI progress’ (for example letters, see Appendix E). Our emails did include links to past ESPAI surveys and included some of our names and affiliations (University of Oxford, AI Impacts, University of Bonn), so evidence about the topic and the authors was available to participants who investigated or were previously familiar with us or the survey (increasingly likely, since our past surveys received substantial public attention, and have now been taken before by many researchers).

As well as bias in who chooses to take a survey, there can be bias in who chooses to answer each particular question within it. All but one question could be skipped. However each question in this survey was answered by on average 96% of those who saw it, excluding demographics questions, free response questions, and questions asking for a response conditionally. So the scope for question-level participation bias in opinion questions reported on in this paper is very small.

The question which was skipped most often (answered by 90.3% of those who saw it) was about the number of years until the occupation of “AI researcher” would be fully automatable, with the fixed-probabilities framing. The question which was skipped the least often was about the long-term value of HLMI (see Section 4.2), answered by 100% of those who saw it, probably because it was not possible to skip and continue the survey.

For random samples of responders (n=369) and of non-responders (n=589), we compared gender, region, PhD start year, number of citations, and work in industry or academia, based on our best guesses from available public data. The most substantial difference between the groups was that those who worked in industry were 61% as likely to respond as the base rate. Women were 66% as likely to respond as the base rate, though were a relatively small portion of respondents or non-respondents overall (around one in ten), which limits how much this could affect the aggregate results. Those with at least 1,000 citations were 69% as likely to respond as the base rate, and people in Asia were 84% as likely as the base rate to respond. We checked for correlations between the collected demographic data and respondents’ beliefs about the arrival or overall outcome of HLMI. Most demographic factors showed minimal influence, but female respondents generally expected less extreme positive or negative outcomes. Additionally, respondents whose undergraduate education was in Asia anticipated an 11 year earlier arrival of High-Level Machine Intelligence (HLMI) than participants from Europe, North America, or other regions combined. See Appendix A for more demographic comparisons.

Another source of evidence comes from people’s reported interest in the topic. One particular concern might be that researchers who have a strong interest in the future of AI might be more likely to participate than those who do not. However, when asked “How much thought have you given in the past to when HLMI (or something similar) will be developed?” only 7.6% said “A great deal.” And when asked “How much thought have you given in the past to social impacts of smarter-than-human machines?” only 10.3% said “A great deal.” This would seem to rule out the possibility that a large proportion of respondents were people for whom the future of AI is an ideological issue.

E Links to Data

Anonymized and cleaned responses: <https://tinyurl.com/espai2023-clean-anon-responses>

Example invitation letter: <https://tinyurl.com/espai2023-invite>

List of task descriptions: <https://tinyurl.com/aitasks>

The number of participants who saw and answered each question: <https://tinyurl.com/espai2023-response-counts>

Further data and analysis as it becomes available: <https://tinyurl.com/espai2023-addenda>

F Yuen’s Test (Bootstrap Version) Results

Variable	P-value	T_y	Trimmed mean difference	Lower bound CI	Upper bound CI
hlmi_50percent	0	4.96	12.49	7.58	17.39
faol_50percent	0.0052	2.84	37.54	13.08	62
task_Rosetta_50percent	0.1106	1.62	2.52	-0.66	5.7
task_translatespeech_50percent	0.0072	2.74	2.7	0.77	4.63
task_translatetext_50percent	0.258	1.25	2.69	-2.25	7.62
task_phonebanking_50percent	0.1862	1.28	1.23	-0.67	3.12
task_groupimages_50percent	0.8058	0.24	0.23	-1.79	2.25
task_oneshotlearning_50percent	0.4778	0.72	1.2	-2.25	4.65
task_3D_50percent	0.0366	2.09	1.9	0.15	3.65
task_transcribe_50percent	0.0116	3.29	2.74	1.06	4.41
task_recording_50percent	0.0748	2.1	1.92	-0.16	3.99
task_provemath_50percent	0.9864	-0.02	-0.08	-7.36	7.2
task_Putnam_50percent	0.3896	0.91	2.92	-4.94	10.77
task_Go_50percent	0.3376	0.95	1.55	-1.59	4.68
task_Starcraft_50percent	0.5176	-0.63	-0.45	-1.81	0.92
task_randomPCgame_50percent	0.21	1.41	2.4	-1.49	6.3
task_AngryBirds_50percent	0.5688	0.55	0.22	-0.64	1.08
task_profAtari_50percent	0.6984	-0.39	-0.44	-2.67	1.8
task_novAtari_50percent	0.1042	1.61	1.6	-0.37	3.56
task_laundry_50percent	0.7116	0.35	0.67	-3.36	4.69
task_run_50percent	0.0362	2.19	3.54	0.29	6.8
task_LEGO_50percent	0.6892	-0.4	-0.44	-2.6	1.72
task_sortlist_50percent	0.5616	0.58	0.85	-2.06	3.75
task_Python_50percent	0.0768	1.95	2.52	-0.65	5.69
task_Googleable_50percent	0.1606	1.55	1.6	-1.09	4.29
task_openGoogleable_50percent	0.0398	2.38	2.25	0.19	4.32
task_NoDefAnswer_50percent	0.0258	2.49	5	1.11	8.9
task_essay_50percent	0.0912	1.69	1.24	-0.22	2.69
task_songTop40_50percent	0.1654	1.41	2.64	-1.24	6.52
task_songartist_50percent	0.0958	1.57	2.38	-0.45	5.22
task_bestseller_50percent	0.0038	3.59	6.06	2.42	9.7
task_explainmove_50percent	0.2466	1.17	1.76	-1.28	4.79
task_poker_50percent	0.3718	0.97	1.04	-1.44	3.52
task_virtualworld_50percent	0.0578	-1.85	-3.53	-7.19	0.12
truckdriver_50percent	0.908	-0.12	-0.14	-2.68	2.4
surgeon_50percent	0.0408	2.06	5.92	0.29	11.55
retail_salesperson_50percent	0.006	2.76	3.48	1.01	5.94
AIresearcher_50percent	0.0066	3.72	21.84	8.63	35.06
final_occupation_50percent	0.0264	2.59	31.73	4.52	58.95