Expert Survey on Progress in AI (ESPAI)
SPAR Final Report

Julius Simonelli and Tim Duffy

julius.simonelli@gmail.com · timfduffy@gmail.com

May 2026

Abstract

The Expert Survey on Progress in AI (ESPAI) has been conducted four times since 2016, polling AI researchers on when they expect key milestones, such as High-Level Machine Intelligence (HLMI) and Full Automation of Labor (FAOL), to be achieved. It also asks how they assess the risks and benefits of advanced AI, including the probability of human extinction or similarly permanent and severe disempowerment. In this final report we present three complementary analyses of the ESPAI data. First, we compare the 2024 survey responses across geographic regions (United States, China, and Europe), finding broadly similar views with modest but informative differences. Second, we track how expert predictions have shifted across the 2016, 2022, 2023, and 2024 waves, documenting a dramatic acceleration in aggregate HLMI timelines and an overall increase in safety concern. Third, we apply clustering methods to the 2024 cleaned responses to describe broad regions of a continuous belief landscape, including a notable "polarized/bimodal" group that assigns high probability to both extremely good and extremely bad outcomes.

Introduction

Understanding how AI researchers think about the trajectory and risks of their field is important for forecasting and for informing policy responses to advanced AI. The Expert Survey on Progress in AI (ESPAI), conducted by AI Impacts, is the largest recurring survey of AI researchers on these questions, with waves in 2016, 2022, 2023, and 2024. Each wave asks respondents to estimate timelines for AI milestones, to assess the probability and severity of risks from advanced AI, including human extinction or similarly permanent and severe disempowerment, and to indicate their views on the importance of safety research.

The 2024 wave received responses from 1,502 AI researchers. In this report, we analyze the 2024 data alongside the three prior waves. Our contributions are:

- **Regional comparisons** of the 2024 respondents, disaggregated by country (US, China, Europe), examining differences in risk perception, outcome expectations, and HLMI timelines.
- **Cross-year trend analysis** across all four survey waves, quantifying how HLMI/FAOL timelines, task-specific predictions, value assessments, extinction/disempowerment risk estimates, and safety attitudes have evolved from 2016 to 2024.
- **Cluster analysis** of the 2024 respondents, using Gaussian Mixture Models to describe broad groupings in belief space and examining how these groups differ on timelines, safety concern, and other variables.

Related Work

This work builds directly on the ESPAI survey series:

- **Grace et al. (2018)** — 2016 wave.
- **Stein-Perlman et al. (2022)** — 2022 wave.
- **Grace et al. (2024)** — 2023 wave.

Complementary survey efforts include the AI Index Report (Maslej et al., 2024), which tracks AI progress along technical and societal dimensions, and Zhang et al. (2022), who surveyed machine learning researchers on ethics and governance of AI.

Methods

Data

We use cleaned and anonymized response data from all four ESPAI waves. For modern waves, the cleaned files retain some unfinished rows, so headline sample sizes use finished responses where available: 2022 (N=531), 2023 (N=2,634), and 2024 (N=1,502). The 2024 cleaned file contains 1,793 rows before this finished-response filter. Each wave asked overlapping but not identical sets of questions, so block randomization in the 2023 and 2024 waves means most respondents answered only a subset of questions, so figures and tables report item-level sample sizes where possible. Respondent location was inferred from IP geolocation at the time the survey was taken; it should not be read as citizenship, affiliation, or home institution. IPs likely to reflect travel, conferences, or VPN artifacts were excluded from regional analyses.

Timeline Aggregation

HLMI and task timelines were elicited in two framings: *year-framing* (respondents provide years until 10%, 50%, and 90% probability) and *fixed-year/probability-framing* (respondents provide probabilities at fixed future horizons). Following Grace et al. (2024), we fit a gamma CDF to each individual's three data points, evaluate those CDFs on a shared year grid, average the probability values at each year, and read percentiles from the aggregate mean CDF. Thus, a reported gamma-CDF "median" is the 50th percentile of the aggregate CDF, not the median respondent's direct 50% answer. The Cross-Year Trends HLMI panel uses direct HLMI blocks only; final-occupation prediction blocks are not pooled into that headline HLMI series.

The regional comparison section uses a simpler approach: the raw median of year-framing respondents' 50%-probability estimates. This yields shorter timelines (e.g., 15 years vs. 18 years for 2024) for two reasons: (1) fixed-year/probability-framing respondents tend to give longer estimates and are excluded from the raw median, and (2) the mean CDF is pulled rightward by long-tail predictions. The two methods answer different summary questions and should not be compared as if they were the same estimand.

Clustering

We apply Gaussian Mixture Models (GMM) with full covariance to describe broad regions of respondent belief space. Because block randomization means most respondents only answered a subset of questions, we run multiple focused analyses on overlapping subsets rather than one clustering requiring all features. Model selection uses BIC with silhouette score as a secondary criterion.

Results

Regional Comparisons

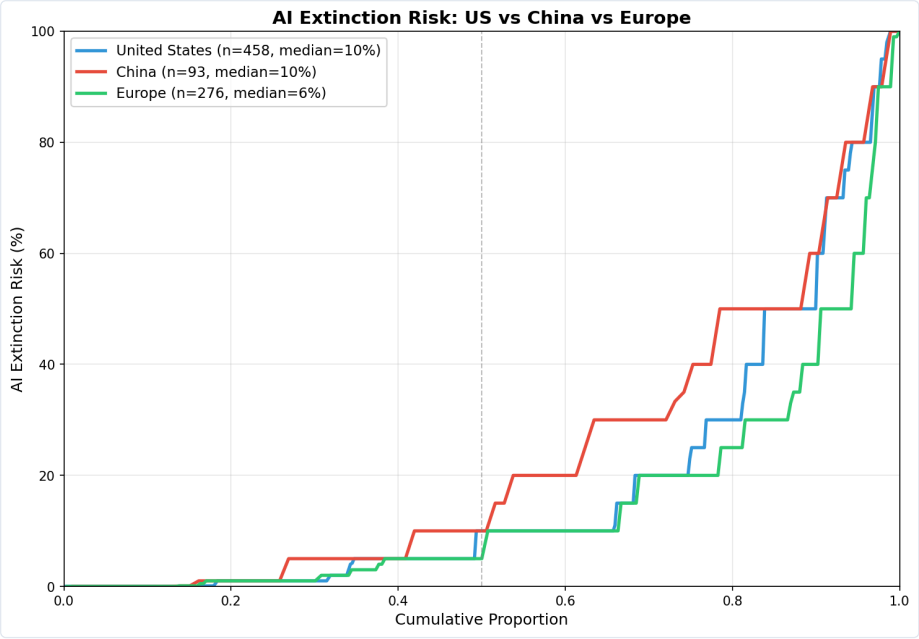
How AI-risk, HLMI-impact, and year-framing HLMI-timeline views compare across the US, China, and Europe (2024 wave)

Questions shown in the regional figures: The extinction/disempowerment CDF uses `extinction_all_1`, the respondent's probability estimate that AI causes human extinction or similarly permanent and severe disempowerment. The HLMI impact bar chart uses `vb_1_1` through `vb_1_5`, the five probability masses for HLMI being extremely good, on balance good, neutral, on balance bad, or extremely bad. The regional HLMI timeline CDFs use the year-framing items `hb_a_1`, `hb_a_2`, and `hb_a_3`, which ask for years until 10%, 50%, and 90% probability of HLMI. The fixed-year/probability-framing items `hb_b_*` are not included in those regional timeline CDFs.

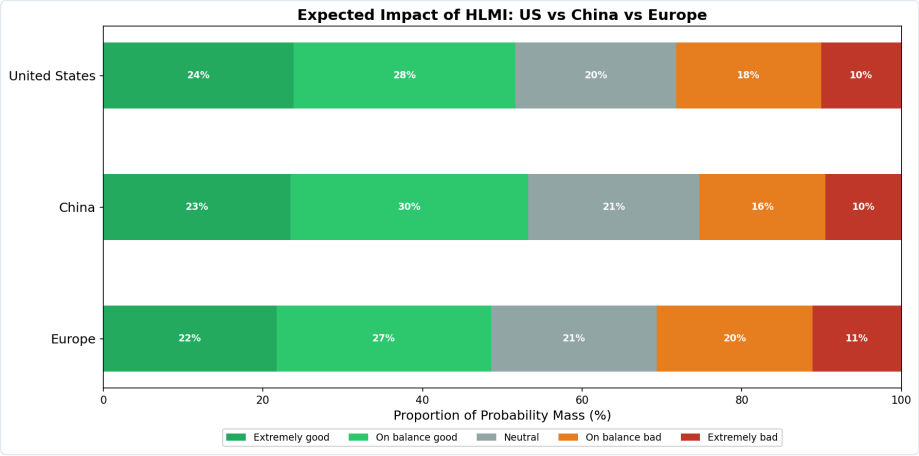
Regional uncertainty note: These regional comparisons are descriptive. The imported regional source provides rendered figures rather than the region-assignment data needed to recompute p-values, effect sizes, or bootstrap intervals in this repo, and the anonymized 2024 clean file does not include regional assignment columns. We therefore describe overlap and tail differences qualitatively rather than adding unsupported significance tests.

Researchers in the US, China, and Europe hold broadly similar views, with modest regional differences

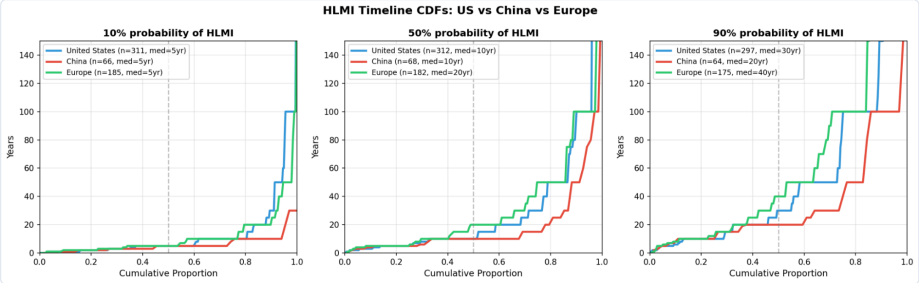
Respondents were classified by the country in which they took the survey[1]; the largest groups were in the US, China, and throughout Europe. Responses across location were fairly similar. On the question about the probability of AI causing human extinction or similarly permanent and severe disempowerment, researchers in China had a somewhat heavier upper tail than their American and European counterparts, though the regional distributions overlap substantially.



Respondents were also asked to evaluate whether High-Level Machine Intelligence (HLMI) would be beneficial or harmful. The three regions returned remarkably similar distributions, with about half of respondents in each assigning probability mass to good or extremely good HLMI outcomes and a quarter to a third assigning probability mass to bad or extremely bad outcomes. Within that narrow band, respondents in China assigned the highest mean probability to positive HLMI outcomes and those in Europe the lowest, while China also had higher estimates on the extinction-or-disempowerment question. This suggests value-of-HLMI assessments and extinction/disempowerment risk estimates are related but not interchangeable. Mean estimates for each outcome are shown below.



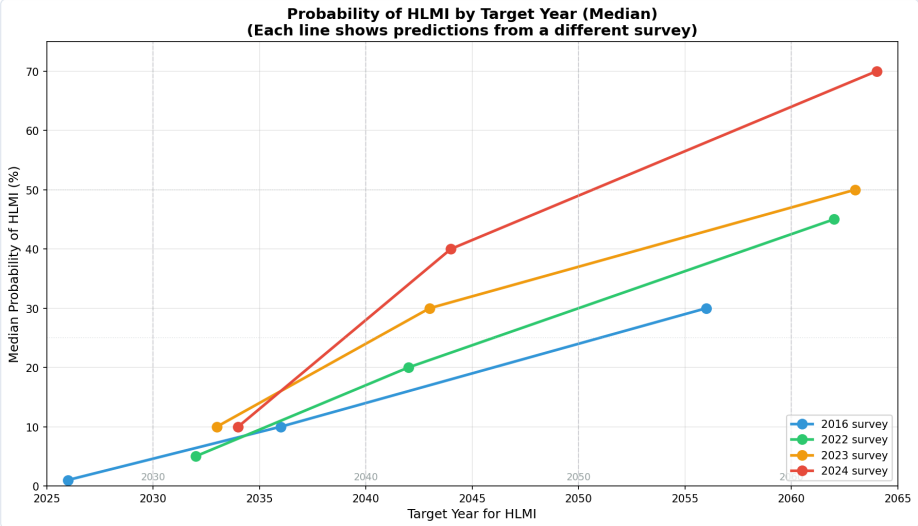
Regarding timelines, the year-framing distributions overlap heavily at shorter timelines, while the upper tail is longer for US and European respondents than for respondents in China. Among researchers with shorter-than-median timelines, estimates were similar across regions. But among those with longer timelines, Americans and Europeans predicted HLMI significantly further out than Chinese researchers. That is, the optimistic ends align, but the cautious ends diverge.



HLMI timelines have shortened in recent survey waves

Expectations for the arrival of HLMI were polled using two methods: estimating the number of years until a 10%, 50%, or 90% probability of arrival, and estimating the probability of arrival within fixed future horizons.

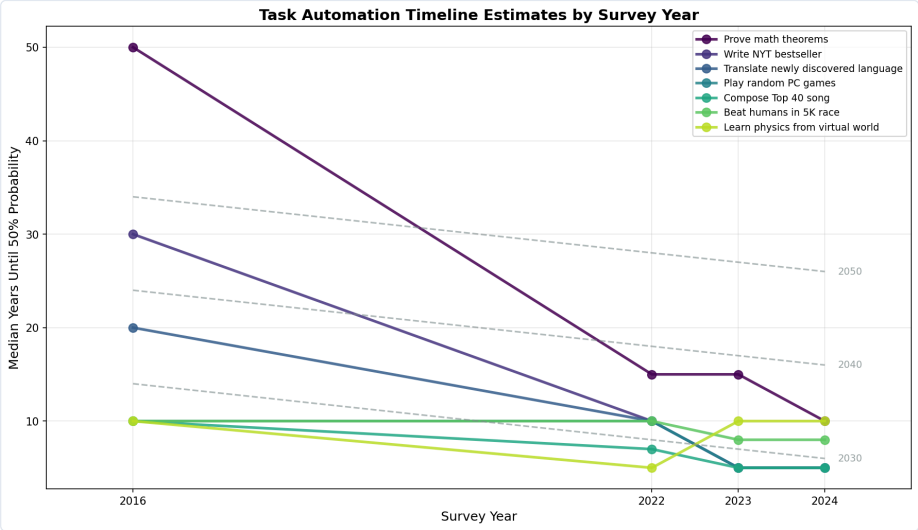
Under both methods, timelines actually extended slightly between 2016 and 2022, but have seen significant shortening each year from 2022 to 2024.



The probability-framing results are shown above. The year-framing question, which asks directly how many years until each probability threshold, yields more aggressive timelines; both methods agree on shortening across recent surveys, and the year-framing analysis appears in the Cross-Year Trends section.

Expected timelines for individual tasks have fallen unevenly

The AI Impacts survey was first conducted in 2016, allowing for an analysis of how task-specific timelines have evolved. For tasks originally expected to take more than a decade, timelines have shortened significantly—particularly in math and language. However, evolution has been mixed for other domains. In many cases, expected dates have actually receded, most notably for robotics-related tasks.



---

[1] Location was determined based on IP address, which leaves some uncertainty since respondents could be using a VPN, or have taken the survey away from home. The timing of the 2024 survey overlapped with NeurIPS 2024; respondents whose IPs resolved to Vancouver during NeurIPS 2024, or to other locations likely to reflect travel, conferences, or VPN artifacts rather than normal survey-taking location, were excluded from regional analyses.

> **A note on methodology:** The HLMI timeline estimates in the Regional Comparisons section are *raw medians* of the year-framing question only—i.e., the median of respondents' direct answers to "how many years until a 50% probability of HLMI?" The Cross-Year Trends section uses gamma CDF fitting (Grace et al., 2024), which combines both year-framing and fixed-year/probability-framing

respondents into a single aggregate distribution and reads off the 50th percentile. The gamma CDF method produces longer timelines (e.g., 18 years vs. 15 years for 2024) because it incorporates the fixed-year/probability-framing group—who tend to give longer estimates—and the mean CDF is pulled rightward by long-tail predictions. The two methods answer different summary questions and should not be compared as if they were the same estimand. See Methods: Timeline Aggregation for details.

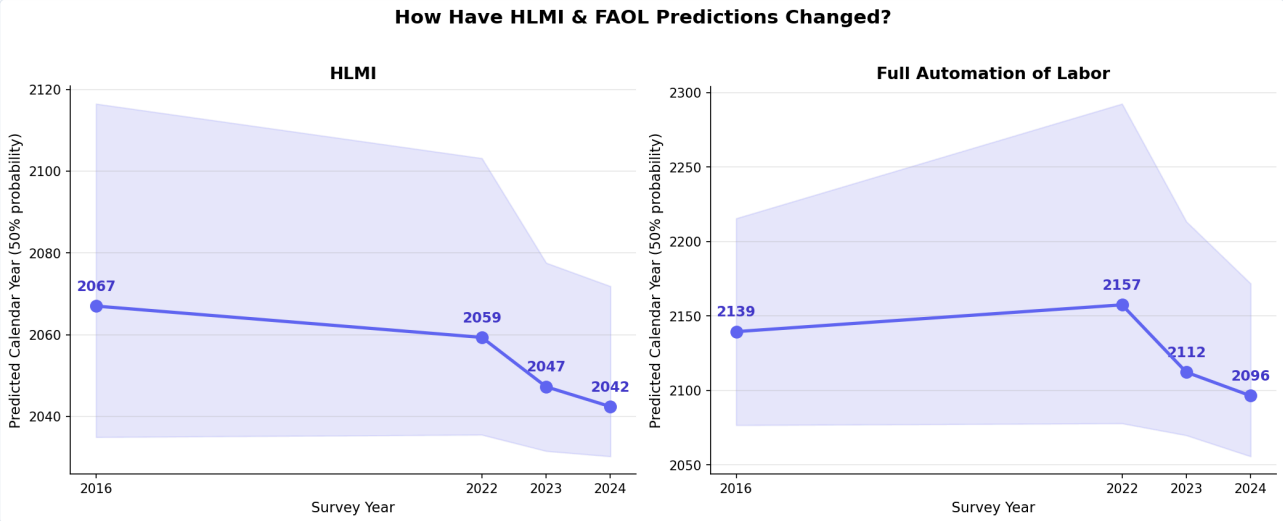
Cross-Year Trends

How expert predictions have shifted across the 2016, 2022, 2023, and 2024 survey waves

1. HLMI & FAOL Timeline Trends

Aggregate-CDF 50th percentile calendar year estimates for High-Level Machine Intelligence and Full Automation of Labor. Computed via gamma CDF fitting on both year-framing and fixed-year/probability-framing respondents.

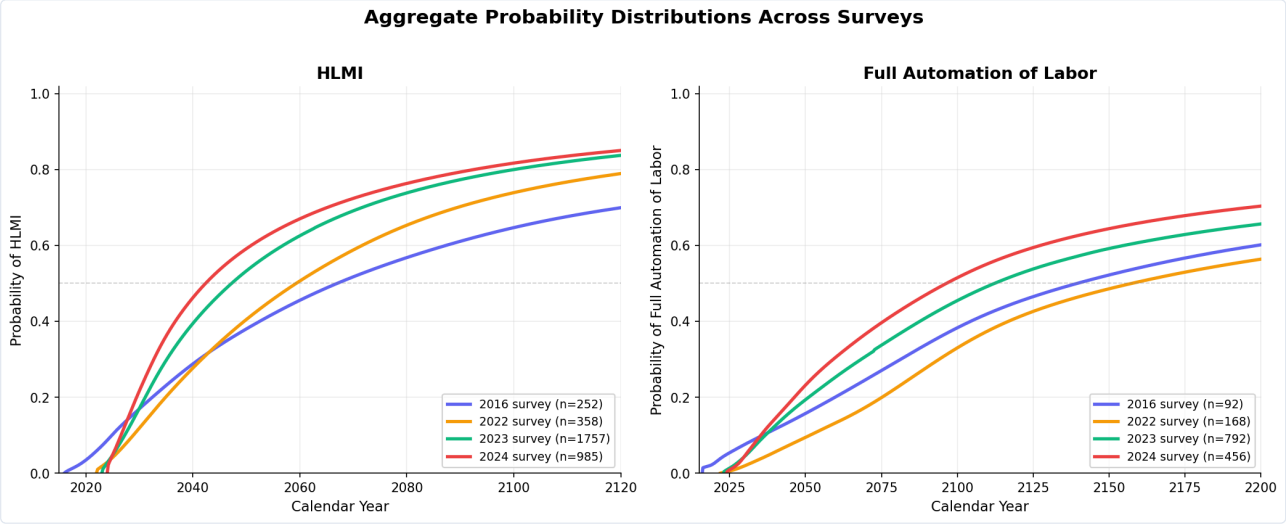
On this measure, direct HLMI estimates moved from 2067 (51.0 years after the 2016 survey) to 2042 (18.4 years after the 2024 survey), a 25-calendar-year earlier estimate. FAOL moved from 2139 to 2096, a 43-calendar-year earlier estimate.



The line is the 50th percentile of the aggregate mean CDF. The shaded band is a robust visual band from the 10th and 90th percentiles of the pointwise median CDF; the mean-CDF 90th percentile is highly sensitive to long right tails and often falls beyond the 500-year percentile cap.

Aggregate probability distributions across surveys

Each curve is the mean mixture CDF: gamma CDFs fitted per respondent, evaluated on a shared year grid, and averaged at each year. The horizontal axis is calendar year; the vertical axis is aggregate probability. Where each curve crosses the dashed 50% line is the 50th percentile of that survey's aggregate CDF. Curves that sit further to the left indicate sooner expectations.



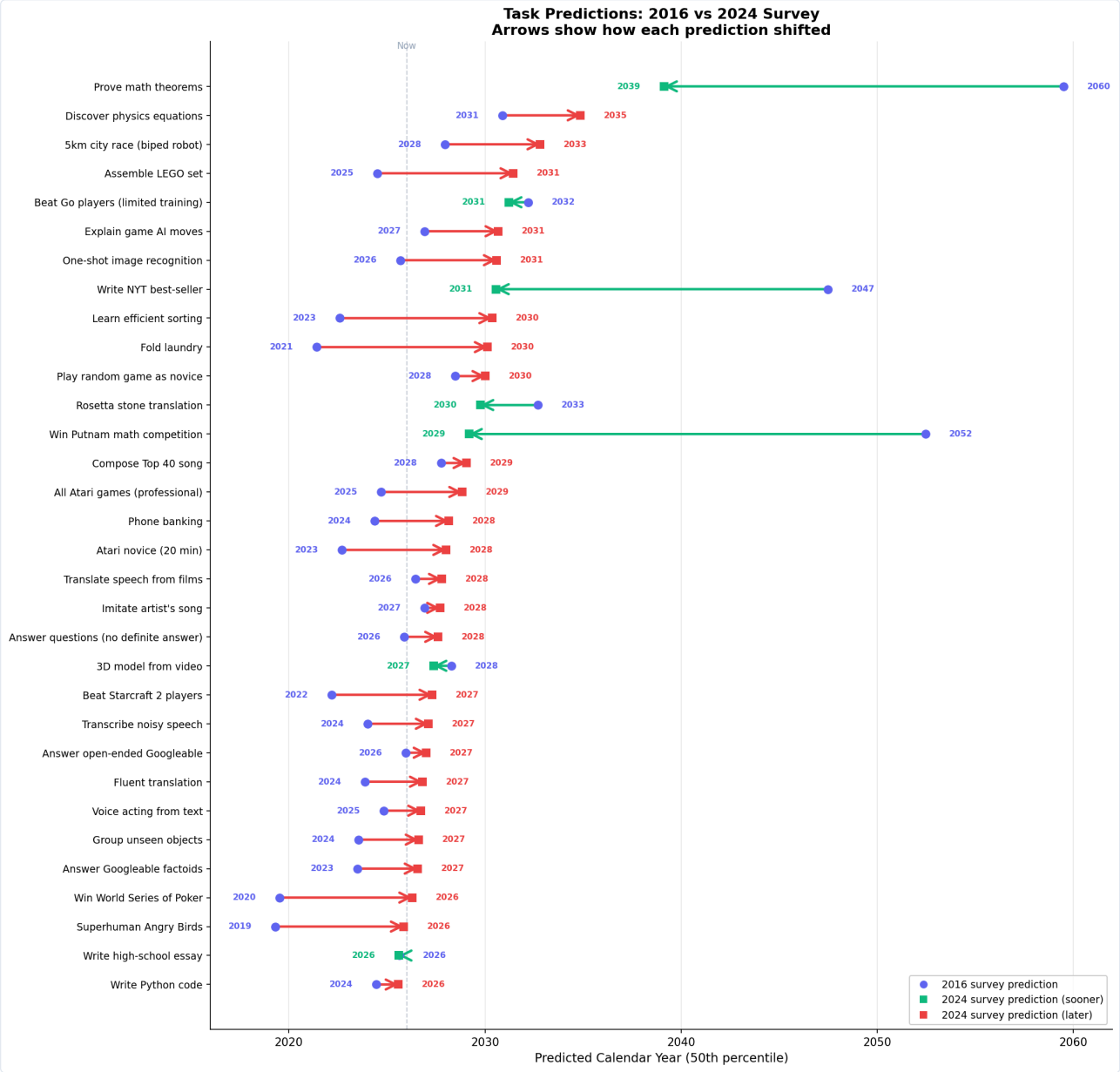
The 2016 survey used somewhat different question wording and was administered to a different population, so the 2016 curve is not strictly apples-to-apples with the 2022/2023/2024 curves and should be read as indicative rather than directly comparable. The 2016 direct-HLMI estimate here is later than the original published 2061 estimate; local validation points to residual 2016 cleaning/respondent-set differences rather than the aggregation code as the likely source.

Survey	HLMI				FAOL			
	Mean-CDF 10th	Mean-CDF 50th	Mean-CDF 90th	N fits	Mean-CDF 10th	Mean-CDF 50th	Mean-CDF 90th	N fits
2016	2025 (8.9 yr)	2067 (51.0 yr)	2504 (488.2 yr)	252	2036 (20.0 yr)	2139 (123.5 yr)	—	92
2022	2029 (6.9 yr)	2059 (37.3 yr)	2284 (261.8 yr)	358	2052 (29.5 yr)	2157 (135.4 yr)	—	168
2023	2027 (4.3 yr)	2047 (24.3 yr)	2197 (174.1 yr)	1757	2037 (13.8 yr)	2112 (89.2 yr)	—	792
2024	2027 (2.6 yr)	2042 (18.4 yr)	2187 (163.3 yr)	985	2035 (11.3 yr)	2096 (72.4 yr)	—	456

2. Task Prediction Trends

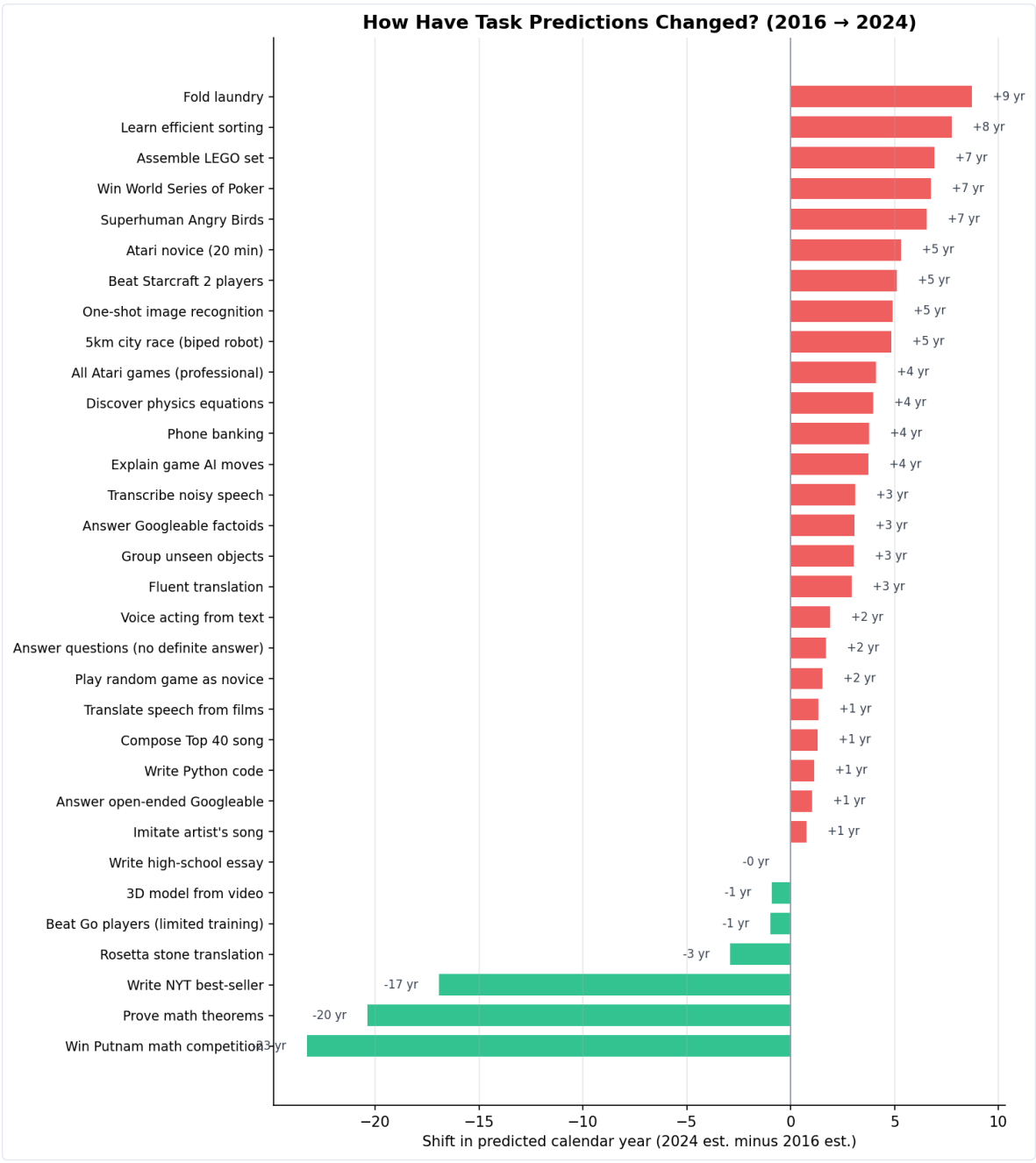
How the aggregate-CDF 50th percentile calendar year for each AI milestone has shifted. Each arrow shows where a task's prediction started (2016 survey, purple dot) and where it ended up (2024 survey, square). Green = prediction moved sooner, red = moved later. Only the 32 tasks present in all survey years are shown.

Reconciliation with the 2023 report graph: Grace et al. (2024) Figure 2 compares 2022 to 2023 forecasts and sorts tasks by their one-year change. This section instead compares 2016 to 2024 for the 32 tasks present in all waves. Both use gamma-CDF aggregation of year-framing and fixed-year/probability-framing responses, but the displayed shift here is a *calendar-year* shift: `(survey year + aggregate-CDF 50th percentile years-from-survey)` in 2024 minus the same quantity in 2016. Thus a task with an unchanged years-from-survey estimate would appear about eight calendar years later simply because the survey date moved from 2016 to 2024.
Task-wording caveat: Some short-horizon task targets may already be feasible, close to feasible, or ambiguous under current systems and the original task wording. Treat very near-term milestones as evidence about respondents' interpretation of the task threshold as well as about technical capability.

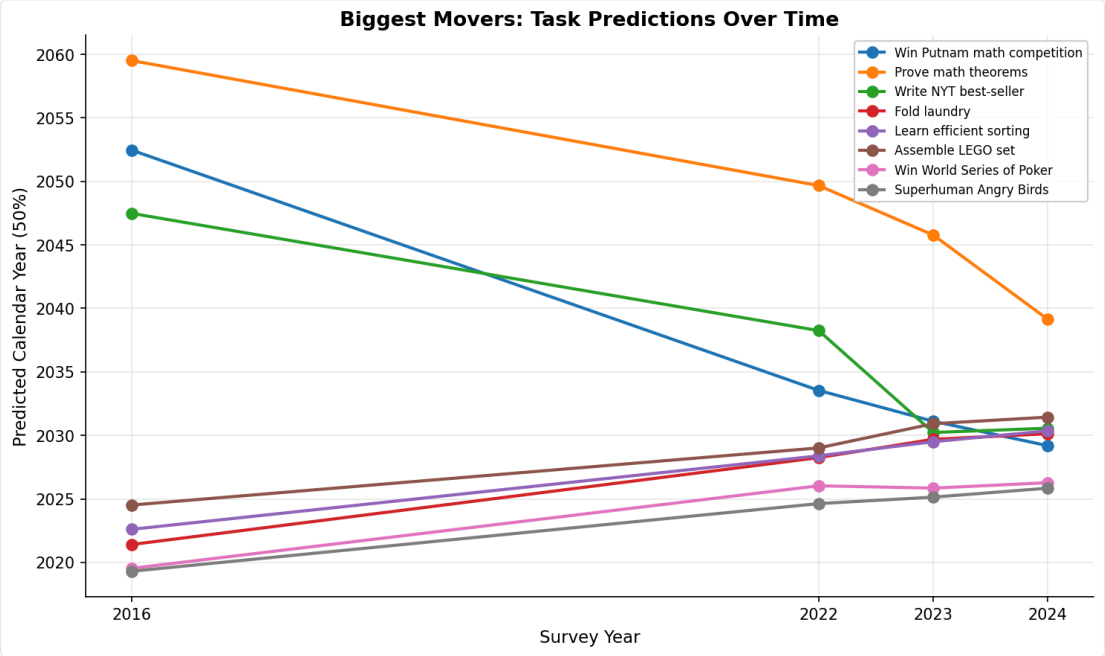


Magnitude of Shift

The same data as above, but focused on the size of the change. Green bars = sooner predictions in 2024 vs 2016 (experts expected faster progress on that task). Red bars = later predictions.



Biggest Movers Over Time



Full Table

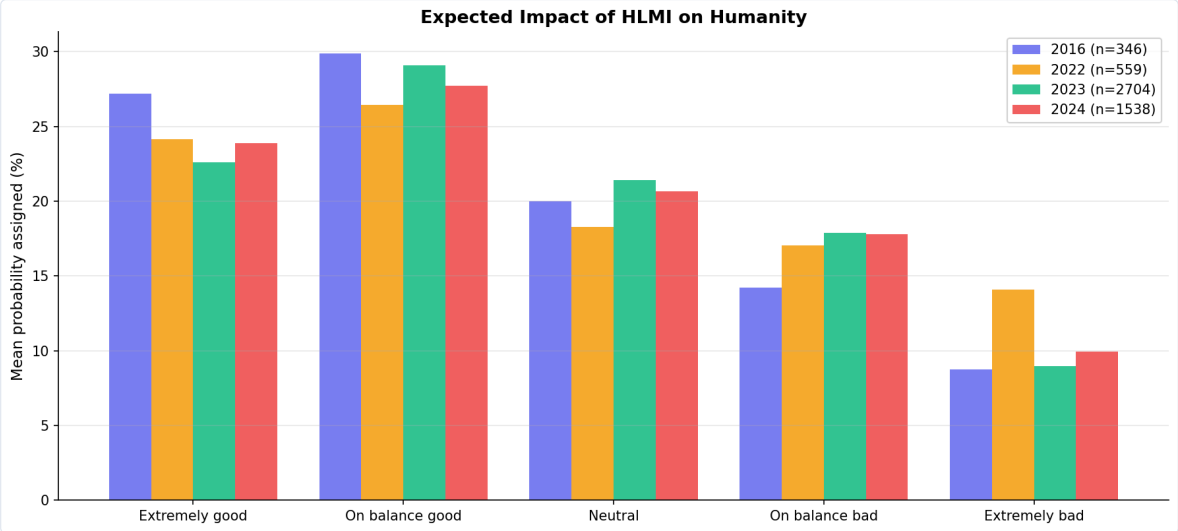
Entries are aggregate-CDF 50th percentile calendar years. Shift is 2024 minus 2016; negative values indicate earlier expected completion in the 2024 survey. These are not raw respondent medians; they are read from pointwise-mean aggregate gamma CDFs fitted to each respondent's task-forecast inputs from whichever framing they received: either three year-framing answers for 10%, 50%, and 90% probability (`ta_*`) or three fixed-year/probability answers at that wave's task horizons (`tb_*`).

Task	2016	2022	2023	2024	Shift
Win Putnam math competition	2052	2034	2031	2029	-23
Prove math theorems	2060	2050	2046	2039	-20
Write NYT best-seller	2047	2038	2030	2031	-17
Rosetta stone translation	2033	2034	2030	2030	-3
Beat Go players (limited training)	2032	2034	2033	2031	-1
3D model from video	2028	2028	2028	2027	-1
Write high-school essay	2026	2025	2025	2026	-0
Imitate artist's song	2027	2028	2027	2028	+1
Answer open-ended Googleable	2026	2028	2026	2027	+1
Write Python code	2024	2027	2025	2026	+1
Compose Top 40 song	2028	2030	2029	2029	+1
Translate speech from films	2026	2029	2028	2028	+1
Play random game as novice	2028	2030	2030	2030	+2
Answer questions (no definite answer)	2026	2030	2027	2028	+2
Voice acting from text	2025	2027	2026	2027	+2
Fluent translation	2024	2029	2026	2027	+3
Group unseen objects	2024	2027	2027	2027	+3

Answer Googleable factoids	2023	2028	2026	2027	+3
Transcribe noisy speech	2024	2027	2026	2027	+3
Explain game AI moves	2027	2032	2031	2031	+4
Phone banking	2024	2029	2028	2028	+4
Discover physics equations	2031	2035	2035	2035	+4
All Atari games (professional)	2025	2027	2029	2029	+4
5km city race (biped robot)	2028	2034	2032	2033	+5
One-shot image recognition	2026	2029	2028	2031	+5
Beat Starcraft 2 players	2022	2025	2027	2027	+5
Atari novice (20 min)	2023	2028	2028	2028	+5
Superhuman Angry Birds	2019	2025	2025	2026	+7
Win World Series of Poker	2020	2026	2026	2026	+7
Assemble LEGO set	2025	2029	2031	2031	+7
Learn efficient sorting	2023	2028	2030	2030	+8
Fold laundry	2021	2028	2030	2030	+9

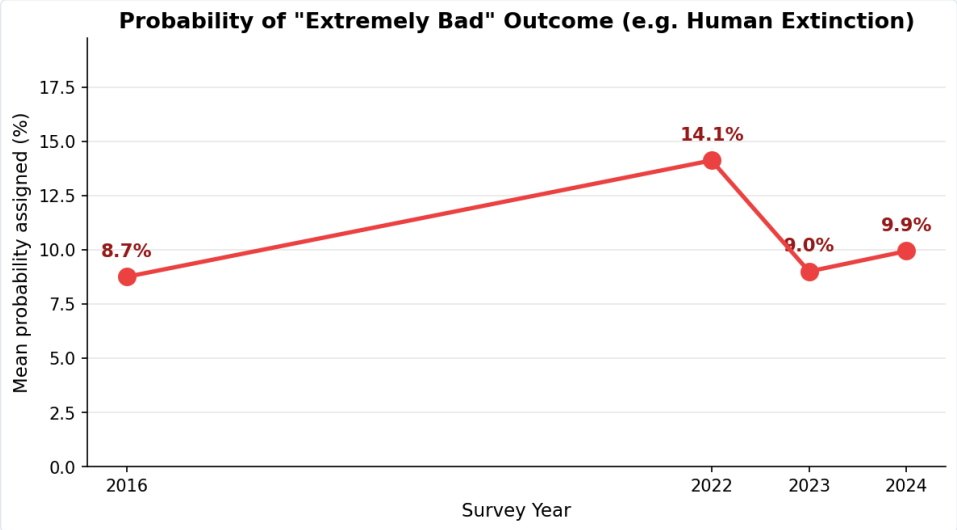
3. Value of HLMI Over Time

Respondents assign probabilities to five impact categories (summing to 100%). Bars show the mean probability assigned to each category.



"Extremely Bad" Outcome Trend

Mean probability assigned to the worst-case value-of-HLMI category, which the survey describes with examples such as human extinction, across survey waves. Note the spike in 2022 followed by a partial retreat.



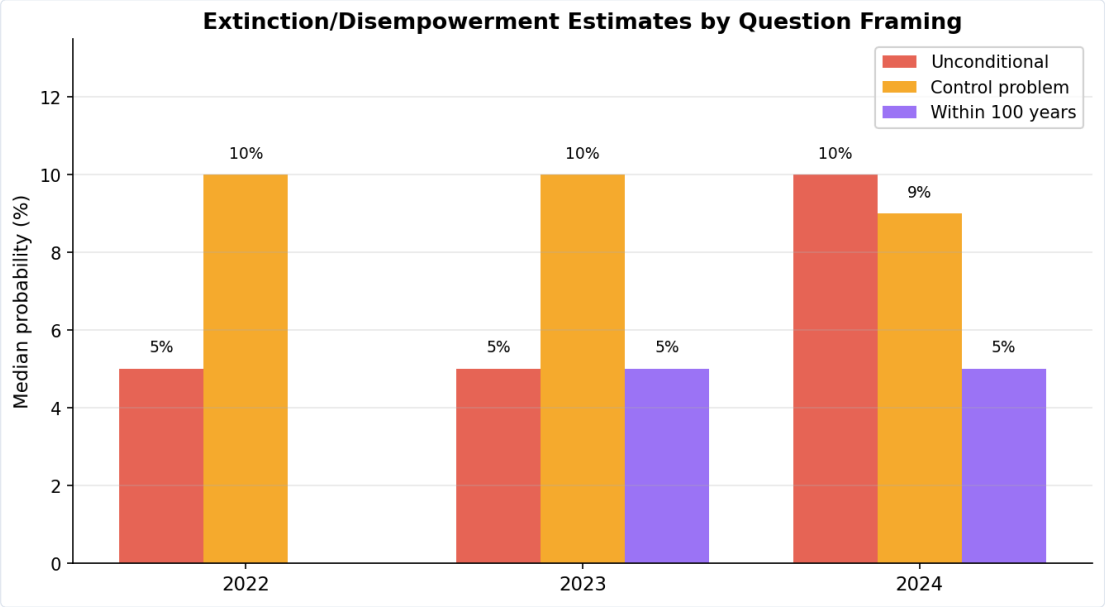
Year	Ext. good	Good	Neutral	Bad	Ext. bad	N
2016	27.2%	29.9%	20.0%	14.2%	8.7%	346
2022	24.1%	26.4%	18.3%	17.0%	14.1%	559
2023	22.6%	29.1%	21.4%	17.9%	9.0%	2704
2024	23.9%	27.7%	20.7%	17.8%	9.9%	1538

4. Extinction/Disempowerment Risk Estimates

Probability of AI causing human extinction or similarly permanent and severe disempowerment, across survey years. Not all questions were asked in all years.

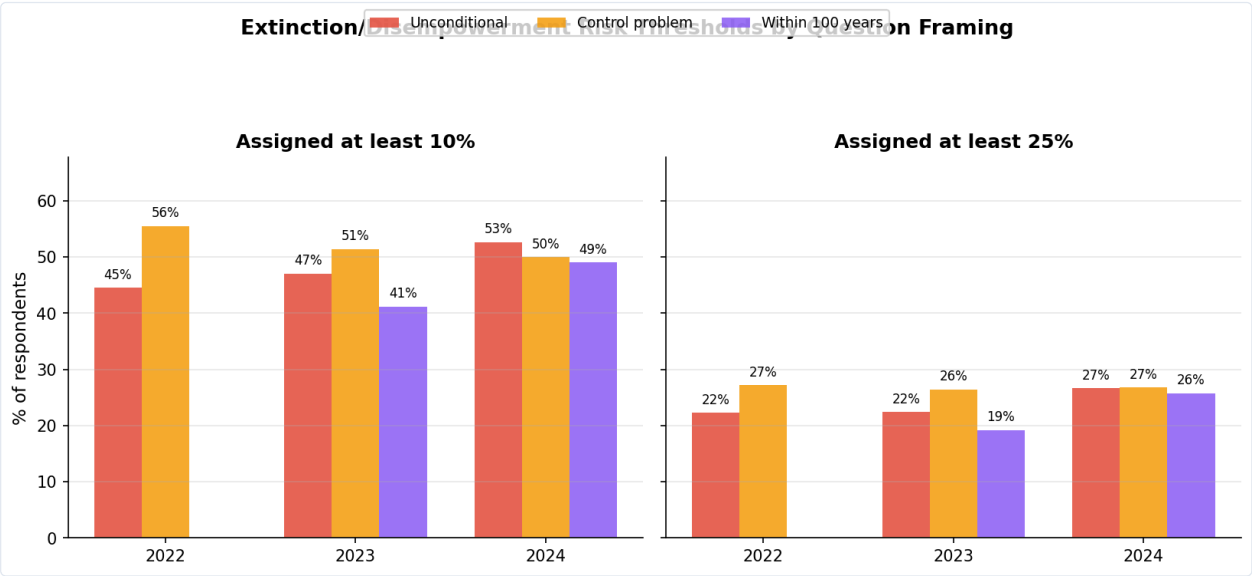
Question-framing variants

Bars compare median estimates across available extinction/disempowerment question framings. The unconditional and control-problem framings are available from 2022 onward; the within-100-years framing is available in 2023 and 2024. These are related but distinct questions, so their estimates should not be treated as one interchangeable series.



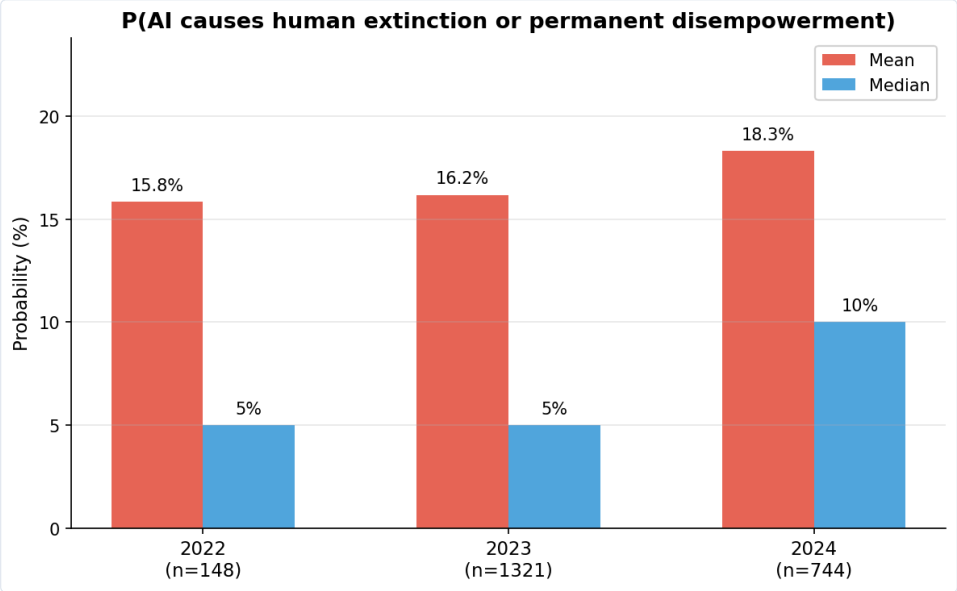
Extinction/Disempowerment Risk Thresholds Over Time

Percentage of respondents assigning at least 10% or 25% probability to AI-caused human extinction or similarly permanent and severe disempowerment, shown separately for each available question framing.



Mean vs Median Extinction/Disempowerment Risk Over Time

Mean and median probability assigned to AI causing human extinction or similarly permanent and severe disempowerment (unconditional question).



Extremely bad outcomes vs extinction/disempowerment

Across overlapping waves (2022-2024), the mean value-of-HLMI probability assigned to "extremely bad" fell from 14.1% to 9.9% after the 2022 spike, while the mean unconditional extinction/disempowerment estimate rose from 15.8% to 18.3% and the median rose from 5.0% to 10.0%. Within the same respondents, the two measures are positively correlated in each modern wave (ρ=0.41 to 0.47), but their aggregate trends are not identical. The value-of-HLMI item is a broad distribution over how good or bad HLMI's long-run effect on humanity would be, while the extinction/disempowerment item asks directly about one severe risk channel.

Year	P(HLMI extremely bad)	P(extinction/disempowerment)	Same-respondent association
2022	14.1% mean, 5.0% median n=559	15.8% mean, 5.0% median n=148	ρ=0.41 paired n=148
2023	9.0% mean, 5.0% median n=2704	16.2% mean, 5.0% median n=1321	ρ=0.47 paired n=1321
2024	9.9% mean, 5.0% median n=1538	18.3% mean, 10.0% median n=744	ρ=0.44 paired n=744

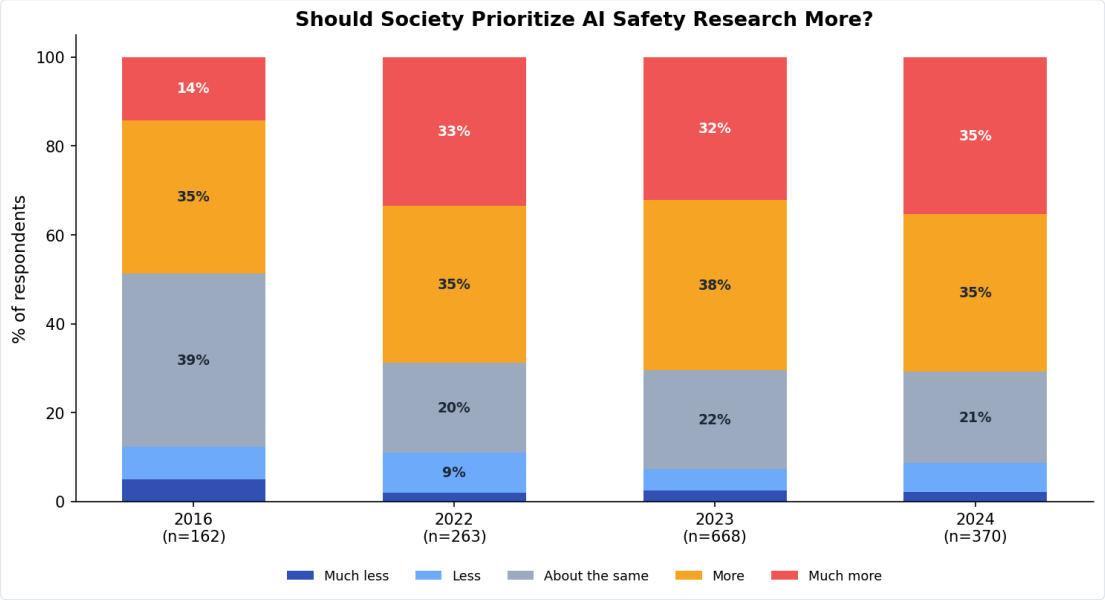
Year	Unconditional (all scenarios)	Due to control problem	Within 100 years
2022	5.0% (mean 15.8%, n=148) ≥10%: 44.6%, ≥25%: 22.3%	10.0% (mean 20.5%, n=162) ≥10%: 55.6%, ≥25%: 27.2%	—
2023	5.0% (mean 16.2%, n=1321) ≥10%: 47.1%, ≥25%: 22.5%	10.0% (mean 19.4%, n=661) ≥10%: 51.4%, ≥25%: 26.5%	5.0% (mean 14.4%, n=655) ≥10%: 41.2%, ≥25%: 19.2%
2024	10.0% (mean 18.3%, n=744) ≥10%: 52.7%, ≥25%: 26.6%	9.0% (mean 18.5%, n=392) ≥10%: 50.0%, ≥25%: 26.8%	5.0% (mean 17.5%, n=353) ≥10%: 49.0%, ≥25%: 25.8%

5. Safety Attitudes Over Time

How researcher views on AI safety have evolved across survey years.

Safety Research Prioritization

"Should society prioritize AI safety research more than it currently does?"



Year	Much less	Less	About the same	More	Much more	N
2016	4.9%	7.4%	38.9%	34.6%	14.2%	162
2022	1.9%	9.1%	20.2%	35.4%	33.5%	263
2023	2.4%	4.9%	22.2%	38.3%	32.2%	668
2024	2.2%	6.5%	20.5%	35.4%	35.4%	370

6. Sample Sizes

Cleaned files may retain unfinished rows. "Finished responses" is the main respondent denominator where a completion indicator exists; question-level analyses use their own item-level sample sizes. HLMI columns below count respondents with at least one usable answer in the year-framing or fixed-year/probability-framing HLMI block. "Task items with data" counts task questions with at least one usable year-framing response, not respondents.

Year	Cleaned rows	Finished responses	HLMI year-framing respondents	HLMI fixed-year respondents	Task items with data
2016	460	322	130	130	32
2022	738	531	179	191	32
2023	3270	2634	912	889	39
2024	1793	1502	537	471	39

Cluster Analysis

Identifying broad respondent groupings in the 2024 survey (N=1,502)

A data-driven exploration of broad groupings and the safety divide among 1,793 cleaned ESPAI 2024 rows.

Executive Summary

1. We tested clustering solutions from K=2 through K=6 and found that four clusters yielded the most informative groupings (n=1,538). Beyond the expected high-value and low-value outlook groups, a **Polarized/Bimodal** group emerges that assigns high probability to both extremely good AND extremely bad outcomes. This is not simply a low-value-outlook or technophobic group -- their P(extremely good) is comparable to the highest-value-outlook cluster. They are better understood as "high-impact believers" who are convinced AI will be transformative but genuinely uncertain whether that transformation will be positive or catastrophic.

2. We observe a meaningful distinction between more and less safety-concerned researchers, though it falls along a spectrum rather than a clean divide. Among 754 respondents with safety data, those in the high-concern group assign 12% mean probability to extremely bad outcomes (vs. 7% for low-concern), and give higher extinction/disempowerment risk estimates (32% vs. 11% mean), with moderate-to-large effect sizes on key measures.

3. Those who see higher extinction/disempowerment risk tend to predict EARLIER HLMI (rho=-0.21, p=0.0000, n=446), suggesting that concern is driven by the belief that powerful AI is coming soon, not that AI is inherently dangerous regardless of timeline.

4. Hardware vs. software beliefs exist on a spectrum, leaning slightly hardware. Among respondents who rated both (n=58), 35 lean hardware and 18 lean software. Computing hardware is rated as the most impactful factor overall (median 60% progress reduction if halved), ahead of algorithms (50%), data (50%), and funding (40%). Notably, people who rate hardware as more important tend to rate safety as LESS important (rho=-0.32, p<0.01), hinting that hardware-focused people may have a more gradualist worldview.

Important caveat: Block randomization means respondents answered different question subsets. Question coverage varies by item. Cross-cutting analyses use smaller samples and should be treated accordingly. Sample sizes are reported throughout.

1. How Many Camps? Broad Value-Outlook Groupings

1.1 Clustering on Value Outlook (n=1,538)

The value-of-HLMI question asked respondents to assign probabilities (summing to 100%) across five outcomes: extremely good, on balance good, neutral, on balance bad, and extremely bad. This question has 86% coverage of cleaned 2024 rows, making it the highest-coverage feature in this analysis. We cluster on these 5 dimensions using Gaussian Mixture Models.

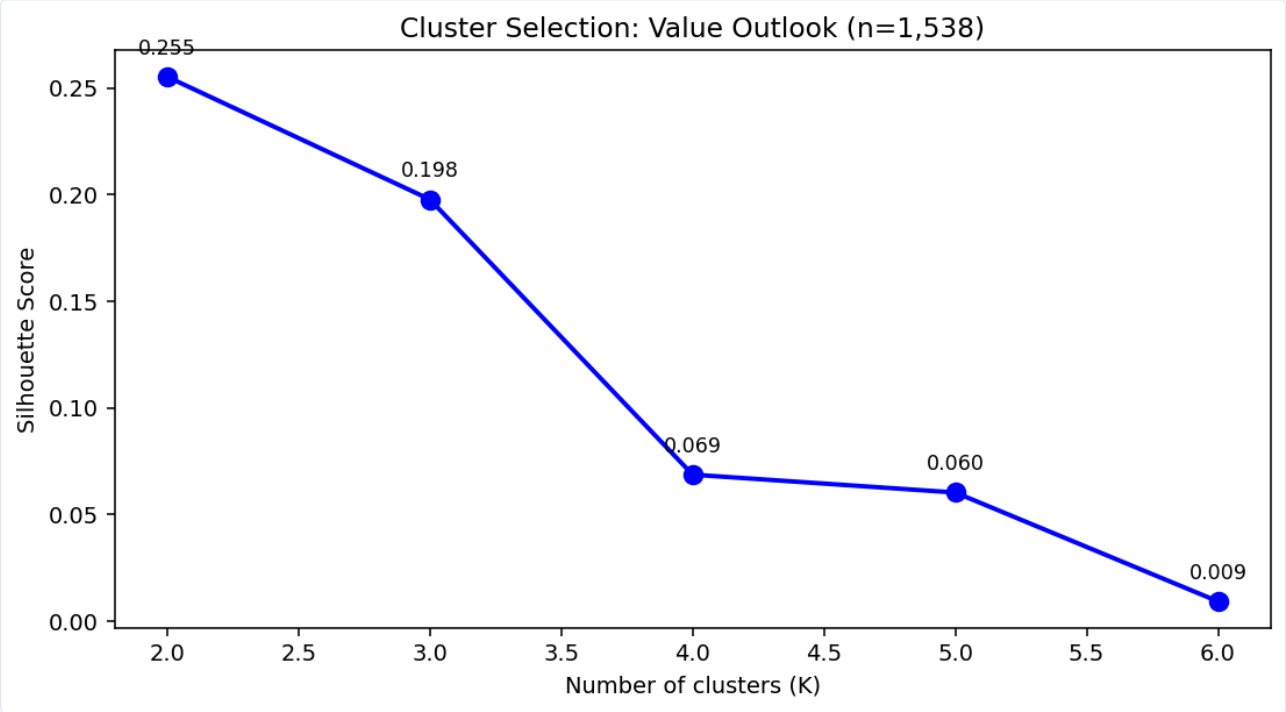


Figure 1: Silhouette scores for K=2 through K=6 clusters on value outlook. K=2 has the highest silhouette, but K=3 and K=4 offer more interpretable structure.

Silhouette scores are modest (0.10-0.12), indicating the groups are not sharply separated. Treat these as broad regions of a continuous opinion landscape, not discrete camps.

1.2 Four-Camp Solution

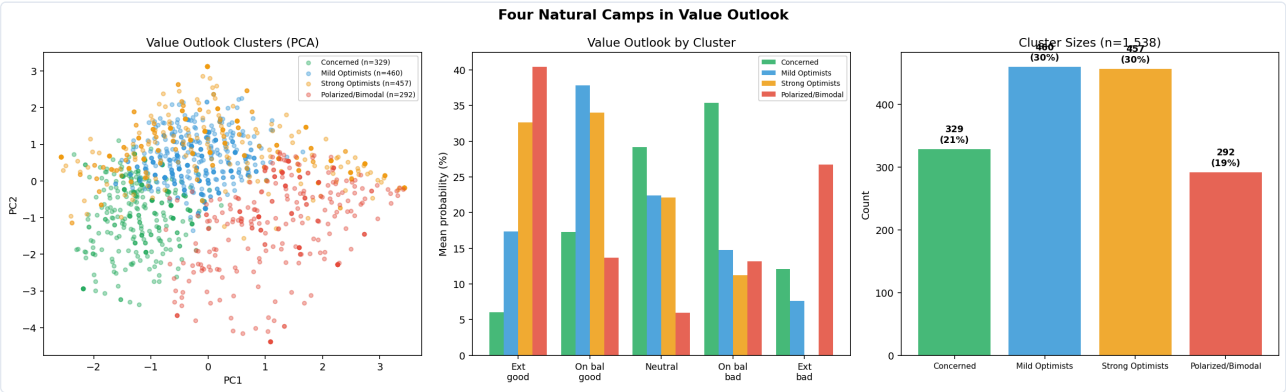


Figure 2: Four natural camps in value outlook. Left: PCA projection. Center: mean probability profiles. Right: cluster sizes.

The most striking finding here is the **Polarized/Bimodal** group. These respondents assign high probability to both extremely good and extremely bad outcomes -- their P(extremely good) is comparable to the Strong Optimists, yet they simultaneously assign substantial probability to catastrophic outcomes. This is not a group of doomers or technophobes. They are researchers who believe AI will be hugely impactful, but who are genuinely torn on whether that impact will be positive or negative. The key axis for this group is *magnitude* of impact, not *direction* -- they have rejected the possibility that AI will be a modest or neutral development.

1.3 Three-Camp Solution (Simpler View)

Collapsing to 3 clusters (which has a higher silhouette score of 0.18 vs 0.07) merges the finer distinctions into a simpler high/moderate/low value-outlook framing. This loses the polarized group but provides a cleaner summary:

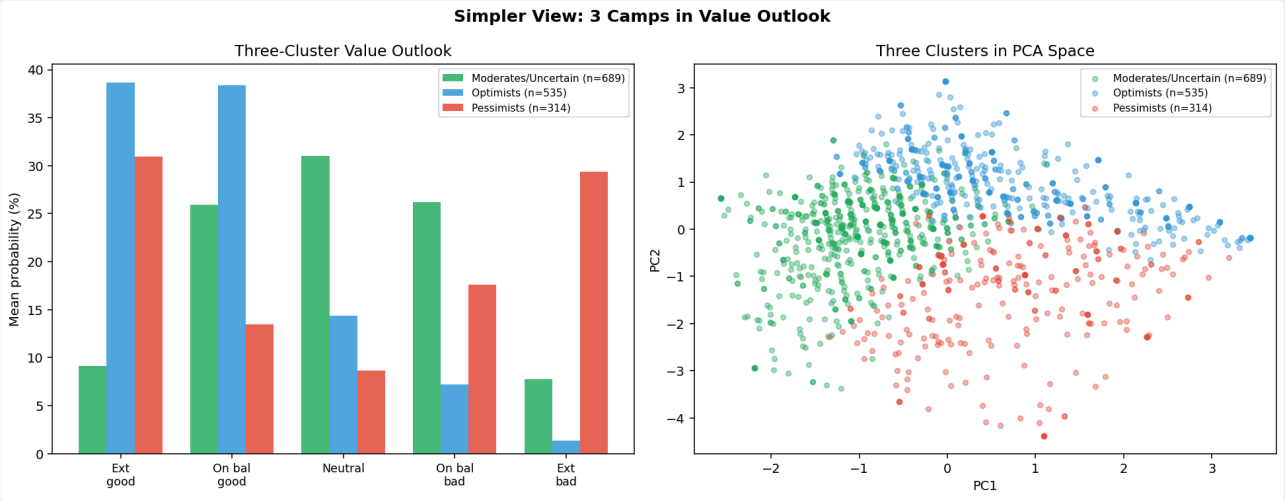


Figure 3: Three-camp simplification. The polarized group is absorbed into the moderate/low-value clusters.

1.4 Intuitive View: How Soon vs. How Good

The PCA axes above lack intuitive meaning. Below, we plot each respondent on two directly interpretable dimensions: their predicted HLMI arrival year (x-axis) and their net optimism score (y-axis), defined as P(good + extremely good) minus P(bad + extremely bad). This reveals where the broad value-outlook regions sit in the space of "how soon" vs. "how beneficial."

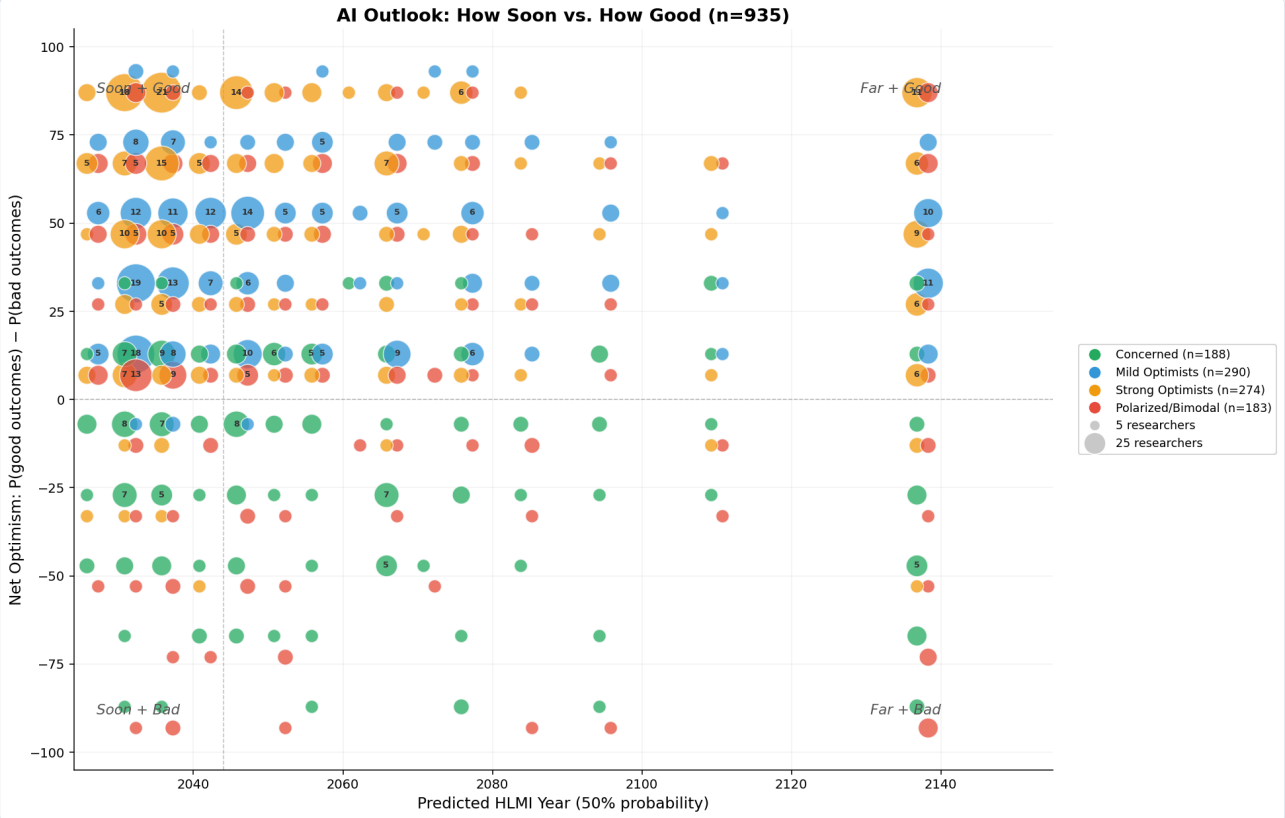


Figure 3b: Respondents (n=935) binned by predicted HLMI year and net optimism. Each bubble's size shows how many researchers from that cluster fall in the bin (labeled when ≥5). The vertical dashed line marks the median predicted year; the horizontal line separates positive net-outlook respondents from negative net-outlook respondents.

2. Safety People vs. Non-Safety People

2.1 Defining the Groups

We built a **safety composite score** from (normalized 0-1 and averaged):

- Safety importance rating (0-4 ordinal scale: not a real problem to among the most important problems in the field)
- Alignment value rating (0-4 ordinal scale: much less valuable to much more valuable than other problems)
- P(Extinction from AI) (0-100%)

Requiring at least 2 of 3 components, we obtained scores for **754** respondents and split at the median (0.500) into High (n=330) and Low (n=424) safety concern groups.

2.2 The Full Comparison

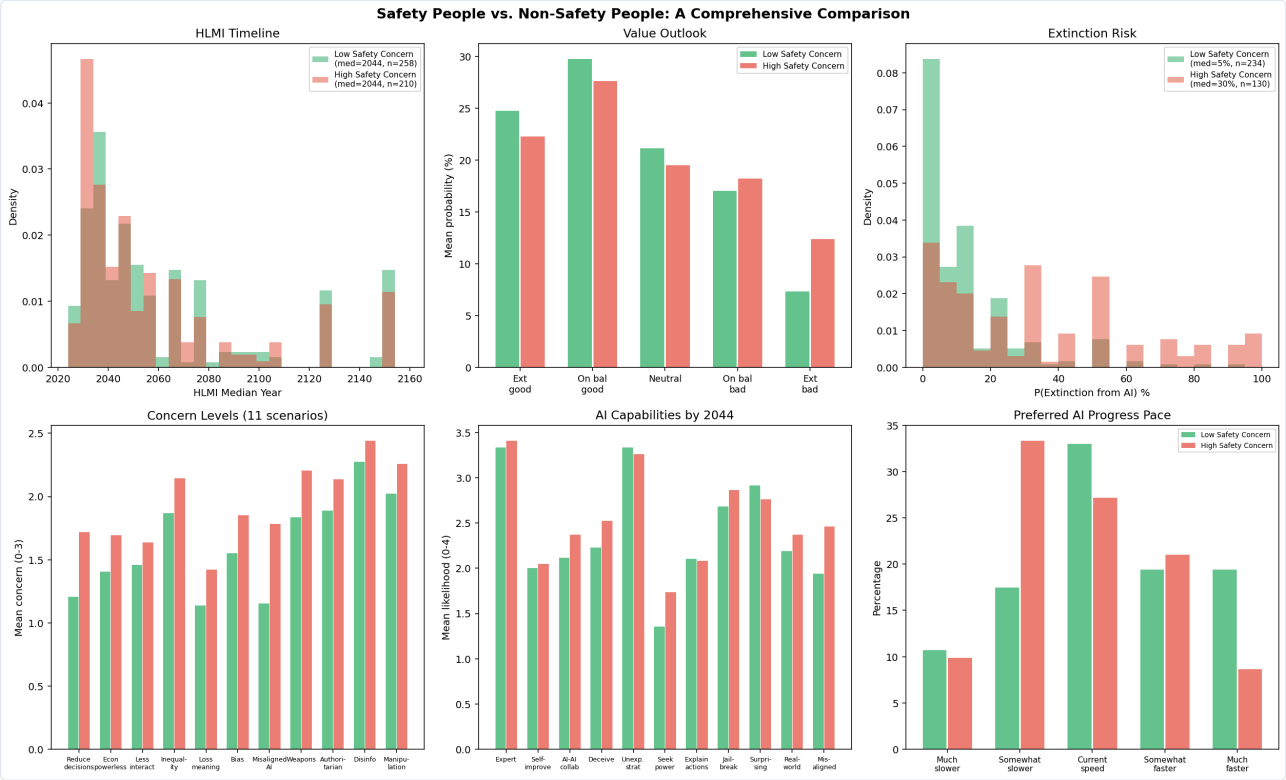


Figure 5: Comprehensive comparison of high-safety vs. low-safety respondents across six dimensions. Safety importance uses the 0-4 ordinal scale above; mean concern averages 11 scenario concern ratings on a 0-3 scale; pace preference is coded 0=much slower, 2=current speed, 4=much faster.

2.3 Statistical Tests

Figure 5 panel	Tested variable	n (High)	n (Low)	Median (High)	Median (Low)	Mean (High)	Mean (Low)	p-value	Effect r
HLMI timeline	HLMI Year	210	258	2044.0	2044.0	2055.4	2060.4	0.0500 n.s.	0.105
Extinction/disempowerment risk	P(extinction/disempowerment)	130	234	30.0	5.0	31.7	10.8	0.0000 ***	-0.466
Value outlook	P(Extremely bad)	330	424	5.0	5.0	12.4	7.3	0.0000 ***	-0.182
Value outlook	P(Extremely good)	330	424	10.0	20.0	22.3	24.8	0.0568 n.s.	0.080
Concern levels	Mean concern	163	220	1.9	1.6	1.9	1.6	0.0000 ***	-0.348
Pace preference	Pace preference	81	103	2.0	2.0	1.9	2.2	0.0501 n.s.	0.164

*Selected Mann-Whitney U tests for scalar summaries from Figure 5. The value-outlook panel is summarized by P(extremely good) and P(extremely bad), the concern panel by mean concern, and the AI-capabilities panel is tested item-by-item in the next table. Effect size r is rank-biserial correlation (|r| > 0.3 = medium, |r| > 0.5 = large). Scale notes: mean concern 0-3, pace preference 0-4, probabilities in percentage points. *** p<0.001, ** p<0.01, * p<0.05*

2.4 AI Capabilities by 2044: What Do Safety People Expect?

Safety-concerned people don't just worry more -- they have systematically different expectations for what AI will be able to do by 2044:

Capability	Hypothesized direction	Mean (High Safety)	Mean (Low Safety)	High - Low	Observed direction	p-value	n
Talk like expert	Exploratory	3.41	3.34	+0.07	High > Low	0.7903 n.s.	199
Self-improve regardless	Exploratory	2.05	2.00	+0.05	High > Low	0.7481 n.s.	199
AI-AI collaborations	High > Low	2.37	2.11	+0.26	High > Low	0.1911 n.s.	199
Deceive humans	High > Low	2.52	2.23	+0.29	High > Low	0.0820 n.s.	197
Unexpected strategies	Exploratory	3.26	3.34	-0.07	High < Low	0.3043 n.s.	200
Seek power	High > Low	1.74	1.35	+0.38	High > Low	0.0104 *	200
Explain actions (trustworthy)	Exploratory	2.08	2.10	-0.02	High < Low	0.7394 n.s.	200
Can be jailbroken	Exploratory	2.86	2.68	+0.18	High > Low	0.3029 n.s.	194
Surprising behavior	Exploratory	2.76	2.91	-0.15	High < Low	0.3611 n.s.	200
Real-world actions	Exploratory	2.37	2.19	+0.18	High > Low	0.2788 n.s.	199
Misaligned goals	High > Low	2.46	1.94	+0.53	High > Low	0.0016 **	199

Scale: 0=Very unlikely, 1=Unlikely, 2=Even chance, 3=Likely, 4=Very likely. High - Low is the high-safety mean minus the low-safety mean, so positive values mean high-safety respondents rated the capability as more likely. Hypothesized direction is marked for risk-relevant capabilities where we expected High > Low; other rows are exploratory.

Overall risk-relevant capability pattern: Averaging the 4 hypothesized High > Low items (AI-AI collaboration, deception, power-seeking, and misaligned goals), high-safety respondents rated these capabilities as more likely by **+0.37** points on the 0-4 scale (means 2.27 vs. 1.90; medians 2.38 vs. 2.00; n=84 high, 115 low; Mann-Whitney p=0.0017 **, |r|=0.26). Respondents are included if they answered at least 2 of the 4 items.

Key pattern: Safety-concerned respondents do rate certain *dangerous capabilities* (power-seeking, AI-AI collaboration, misaligned goals, deception) as more likely, while agreeing with others on benign capabilities (expert conversation, explaining actions). However, the magnitude of these differences is notably small -- typically 0.3-0.6 points on a 0-4 scale. Both groups largely agree on *what AI will be able to do* by 2044. This suggests the safety divide is driven less by disagreements about AI's technical capabilities and more by differences in values, risk tolerance, or beliefs about how those capabilities will be managed.

2.5 Intelligence Explosion Beliefs

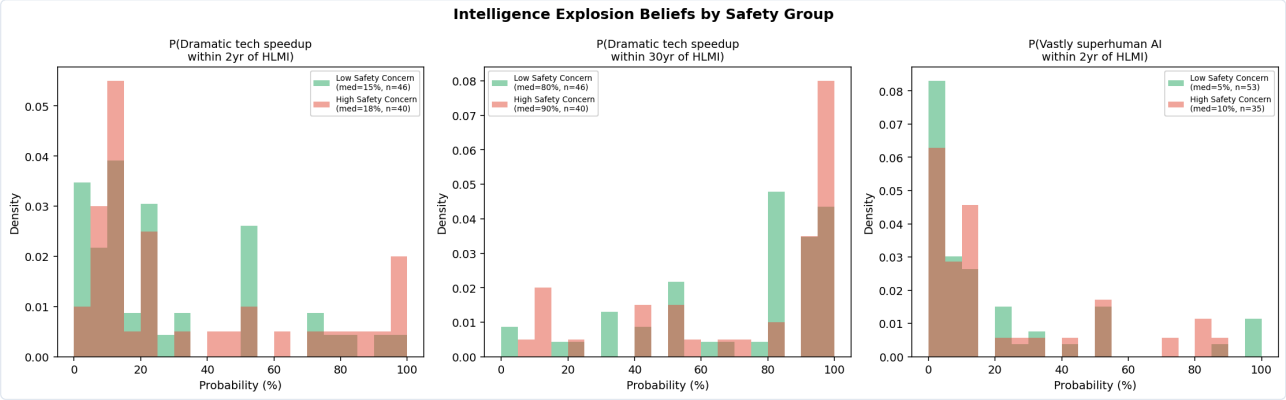


Figure 6: Probability estimates for intelligence explosion scenarios, by safety concern level.

3. How Outlook Predicts Everything Else

The three value-outlook clusters (from Section 1) predict views across many other dimensions:

Variable	Concerned	Mild Optimists	Strong Optimists	Polarized/Bimodal
HLMI Year	2046 [2034-2074]; mean 2060 (n=188)	2044 [2032-2064]; mean 2055 (n=290)	2044 [2034-2064]; mean 2059 (n=274)	2044 [2033-2069]; mean 2058 (n=183)
P(extinction/disempowerment)	10.0 (n=166)	7.5 (n=210)	1.0 (n=223)	20.0 (n=145)
Safety importance	3.0 (n=162)	3.0 (n=220)	2.0 (n=228)	3.0 (n=145)
Mean concern	2.0 (n=165)	1.7 (n=243)	1.5 (n=198)	1.8 (n=148)
Pace preference	1.0 (n=75)	2.0 (n=111)	2.0 (n=109)	2.0 (n=77)

Values are medians with sample sizes in parentheses, except HLMI Year, which shows median [IQR], mean, and item-level n because several groups share a 2044 median. HLMI Year is an interpolated/capped individual 50% calendar-year estimate, not the aggregate gamma-CDF median used in the cross-year section. Scale notes: safety importance 0-4, mean concern 0-3, pace preference 0=much slower to 4=much faster, probabilities in percentage points.

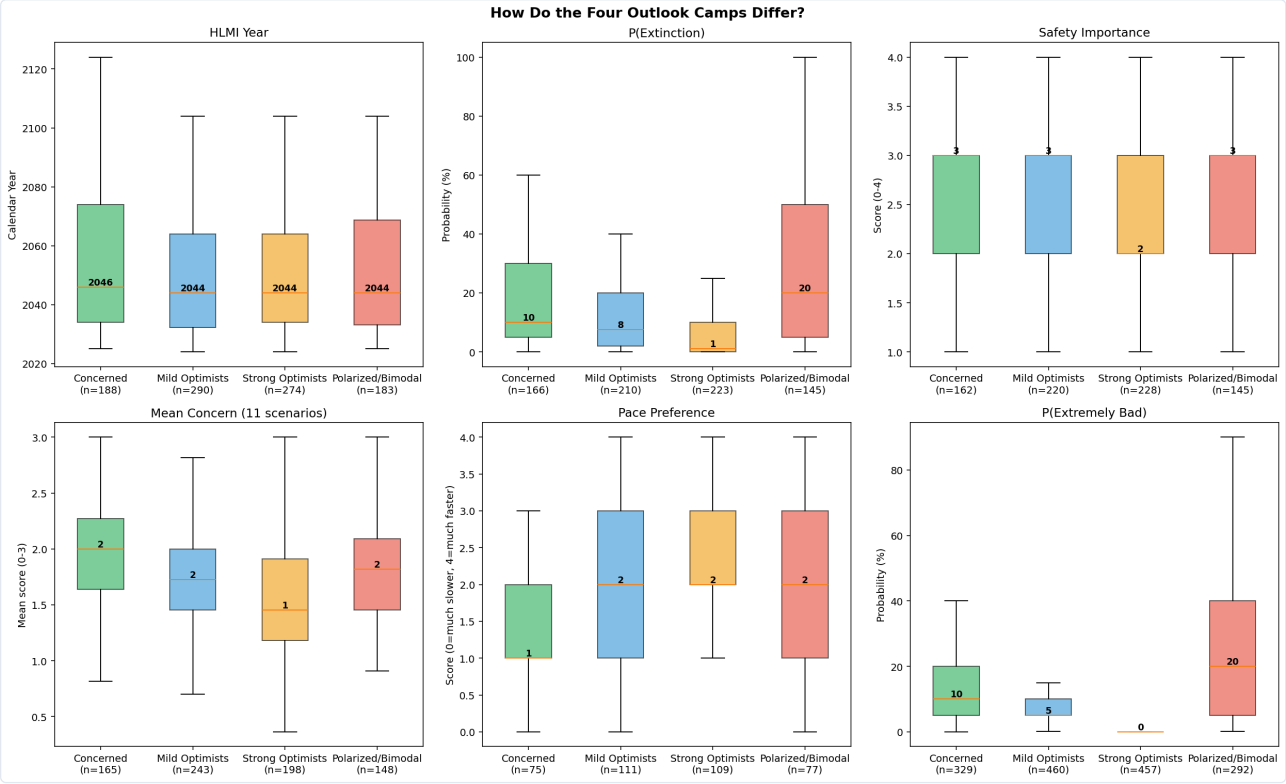


Figure 8: How the four value-outlook camps compare across timelines, extinction/disempowerment risk, safety views, concern levels, pace preference, and P(extremely bad). Mean concern averages 11 scenario concern ratings on a 0-3 scale; safety importance and pace preference use 0-4 ordinal scales.

4. The Full Correlation Structure

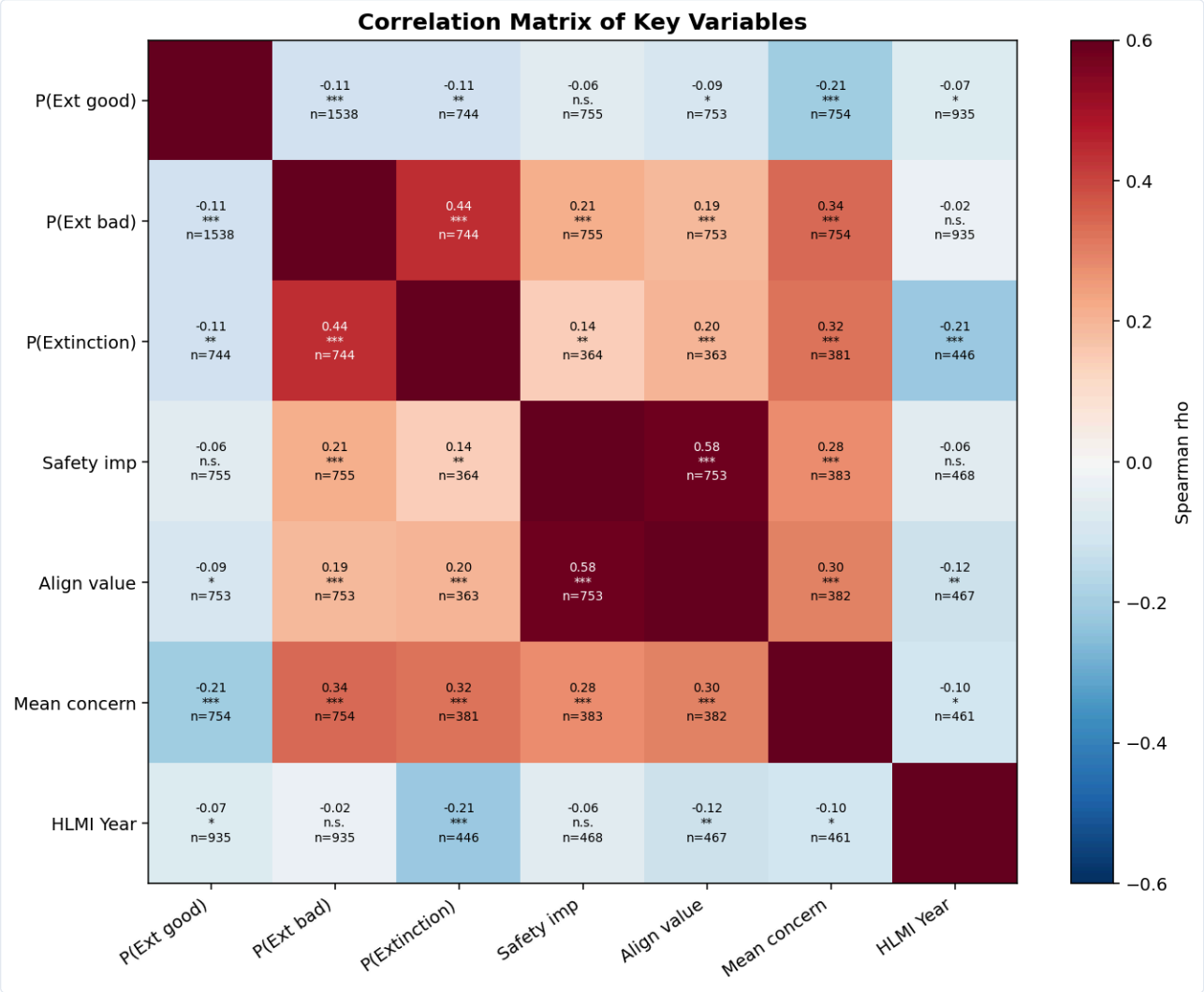


Figure 10: Spearman rank correlations between key variables. Stars indicate significance. Sample sizes shown in each cell. Safety importance uses a 0-4 ordinal scale from not a real problem to among the field's most important problems; alignment value uses a 0-4 ordinal scale from much less valuable to much more valuable than other problems; mean concern uses a 0-3 scale; HLMI Year is a calendar-year estimate; probabilities are 0-100 percentages.

Key Pairwise Correlations

Relationship	Spearman rho	p-value	n
P(extinction/disempowerment) x HLMI Year	-0.215	0.0000 ***	446
P(extinction/disempowerment) x P(Ext bad)	0.436	0.0000 ***	744
Safety imp x P(extinction/disempowerment)	0.141	0.0072 **	364
Safety imp x HLMI Year	-0.056	0.2268 n.s.	468
Safety imp x Mean concern	0.276	0.0000 ***	383
P(Ext good) x P(Ext bad)	-0.113	0.0000 ***	1538
HLMI Year x P(Ext bad)	-0.024	0.4668 n.s.	935
HLMI Year x P(Ext good)	-0.075	0.0226 *	935
Mean concern x P(extinction/disempowerment)	0.321	0.0000 ***	381
Mean concern x HLMI Year	-0.101	0.0300 *	461

Relationship	Spearman rho	p-value	n
Pace pref x P(extinction/disempowerment)	-0.158	0.0327 *	182
Pace pref x HLMI Year	-0.211	0.0010 ***	242
Pace pref x Safety imp	-0.016	0.8289 n.s.	184

Strongest relationships:

- **Safety importance and alignment value** are highly correlated (these aren't independent beliefs -- they form a coherent "safety worldview")
- **P(extinction/disempowerment) and P(Extremely bad)** are strongly linked -- people who assign high extinction/disempowerment risk also see HLMI as likely to be extremely bad
- **P(extinction/disempowerment) and HLMI timeline** are negatively correlated -- people with the highest extinction/disempowerment estimates tend to predict earlier arrival of HLMI, not later
- **Pace preference and HLMI timeline** are negatively correlated -- people with shorter timelines prefer slower progress

Discussion

Several findings stand out across the three analyses. First, HLMI timelines have shortened dramatically—the direct-HLMI aggregate-CDF 50th percentile moved from 2067 (51 years after the 2016 survey) to 2042 (18 years after the 2024 survey)—and this acceleration is consistent across both elicitation framings, though the magnitude differs. Task-level predictions have moved unevenly: mathematical and language tasks have seen the largest acceleration (consistent with rapid progress in LLMs), while robotics tasks have actually receded, suggesting experts view embodied AI as a distinct and harder challenge.

Second, safety concern has risen steadily. The share of respondents favoring more safety research increased from 49% (2016) to 71% (2024), and the median unconditional extinction/disempowerment risk estimate has risen from 5% (2022) to 10% (2024), where this question asks about human extinction or similarly permanent and severe disempowerment caused by AI. The broader value-of-HLMI "extremely bad" item moved differently: it spiked in 2022 and then partially retreated by 2024, even though it remains positively correlated with extinction/disempowerment estimates within each modern wave. Notably, the cluster analysis reveals that higher extinction/disempowerment risk estimates are correlated with *shorter* HLMI timelines ($\rho = -0.21$), suggesting that the most concerned researchers are those who believe powerful AI is arriving soon—not those with abstract, long-horizon worries.

Third, regional differences are more modest than is sometimes assumed. US and Chinese researchers hold broadly similar views on timelines and risks, with Chinese researchers assigning higher mean probability to positive HLMI outcomes while also having a heavier upper tail on the extinction/disempowerment question. European researchers assign the lowest mean probability to positive HLMI outcomes, but are not the highest on extinction/disempowerment risk, suggesting that value-of-HLMI assessments and extinction/disempowerment risk estimates are related but partially independent attitudes.

Limitations

Our analyses have several limitations. The cluster analysis silhouette scores (0.08–0.12) indicate soft, overlapping groupings rather than discrete camps; the four-group framing is an interpretive convenience. Block randomization means not all respondents answered all questions, limiting cross-cutting analyses. We do not control for demographics (career stage, subfield, institution) that may confound clustering. The methodological difference between raw medians and gamma CDF estimates, while documented, means that readers must attend to which method underlies a given figure.

Conclusion

This report presents three complementary views of expert opinion on AI progress and risk, drawing on four waves of the ESPAI survey (2016–2024). We document a pronounced acceleration in HLMI timeline expectations, an overall increase in safety concern, and a belief landscape that can be categorized into four broad, overlapping regions whose views on timelines, risk, and safety tend to differ systematically. Regional variation exists but is modest relative to the overall distribution. These findings suggest that the AI research community is converging toward shorter timelines and greater concern. We believe this shift is relevant for policymakers, funders, and safety researchers planning on longer horizons.

References

1. Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). "When Will AI Exceed Human Performance? Evidence from AI Experts." *Journal of Artificial Intelligence Research*, 62, 729–754.
2. Grace, K., Stewart, H., Sandbrink, J. B., Thomas, S., Weinstein-Raun, B., & Brauner, J. (2024). "Thousands of AI Authors on the Future of AI." arXiv:2401.02843.
3. Maslej, N., et al. (2024). *The AI Index 2024 Annual Report.* Stanford Institute for Human-Centered AI.
4. Stein-Perlman, Z., Weinstein-Raun, B., & Grace, K. (2022). "2022 Expert Survey on Progress in AI." AI Impacts.
5. Zhang, B., Anderljung, M., Kahn, L., Dreksler, N., Horowitz, M. C., & Dafoe, A. (2022). "Ethics and Governance of Artificial Intelligence: Evidence from a Survey of Machine Learning Researchers." *Journal of Artificial Intelligence Research*, 71, 591–666.