EMBARGOED UNTIL 0:01 AM PT, THURSDAY JANUARY 4, 2024

Survey: Median AI expert says 5% chance of human extinction from AI

BERKELEY, CALIFORNIA: In a new survey of 2,778 AI experts, experts gave a median 5% chance that AI would cause human extinction.

In the survey conducted by AI Impacts, a Berkeley-based think tank, in collaboration with researchers at the University of Bonn and University of Oxford, experts offered a median response of 5% for the chance of future AI advances causing human extinction or similarly permanent and severe disempowerment of the human species. Mean responses indicated an even higher risk, suggesting a nearly one-in-six (16%) chance of catastrophic outcomes — the same odds as dying in a game of Russian roulette.

All respondents had recently published a paper in one of six top peer-reviewed AI venues, demonstrating that this is a prevailing view among experts about the potential risks of an emerging technology, the inner workings of which are poorly understood. One in ten respondents put at least a 25% chance on extremely bad outcomes like human extinction, with 1% of respondents thinking there was at least a 75% chance of such outcomes.

Experts also expressed particular concern over the spread of disinformation (including deepfakes), the potential of AI to exacerbate authoritarianism, and its ability to assist dangerous groups in making powerful tools like engineered viruses.

The survey also found that powerful AI may arrive much sooner than many people expect. Experts now see a 50% chance of AI models outperforming humans in every task by 2047, assuming no major disruption to scientific activity — just over 20 years from now. Notably, this estimated date is 13 years sooner than the date arrived at by a similar survey in 2022, in which respondents estimated such AI would not arrive until 2060.

When asked a similar question — when all occupations would be fully automatable — experts had a much longer time horizon, saying there was a 50% chance of this happening by 2116 — 92 years from now. This also represents an advance of 48 years on their 2022 prediction, however: in 2022, experts thought this wouldn’t happen until 2164.

Experts expect systems will be able to do many tasks long before then, though. They thought there was a 50% chance that AI models would be able to produce a new song indistinguishable from a Taylor Swift song by 2027, and a 50% chance AI could write NYT best-selling fiction by 2031. They also predicted a 50% chance that AI models would be able to perform as well as the best humans in the prestigious Putnam math competition within eight years.

In response to the rapid advance of potentially dangerous AI, 70% of experts thought that AI safety should be prioritised more than it currently is, with 36% saying it should be prioritised "more" and a further 34% saying it should be prioritised "much more."

Commenting on the results, **Katja Grace**, lead researcher at AI Impacts, said: "These results show that AI experts think powerful AI may pose substantial risks to humanity, and that it is likely coming sooner than has been anticipated. Beyond this, the confluence of other problems researchers consider concerning deserves its own alarm — our generation's task may be not just navigating a potentially world-ending technology, but doing so while public discussion is distorted, everyone's information is compromised, tyrants and bioterrorists are gaining power, and fresh forces for inequality and injustice are eating at society."

Grace added: "Researchers in academia and industry are on the same page about the risk, and there is broad agreement that society should be doing more research to make AI safer. I think we can get ahead of these risks if we are serious about tackling them now. This is our chance to get AI right."

Information on the survey

The 2023 Expert Survey on Progress in AI was conducted in October 2023. The survey was taken by 2,778 AI experts who had published in the last year in top peer-reviewed venues (NeurIPS, ICML, ICLR, AAAI, IJCAI, JMLR). To keep the survey brief, at several points each participant received questions on only one of several topics, allocated randomly. The researchers allocated these questions to differently sized portions of participants based on factors like the importance of the question and the value of a larger sample size. This means that most questions were not assigned to all 2,778 participants.

Details on the results highlighted in this press release are included below.

For more information on the survey and its methodology, see the paper here: https://aiimpacts.org/wp-content/uploads/2023/04/Thousands_of_AI_authors_on_the_future_of_AI.pdf

For further inquiries or to arrange interviews with Katja Grace, please contact shakeel@aiscc.org.

Selected results

Likelihood of human extinction

Question	N	Median	Mean
What probability do you put on future AI advances causing human extinction or similarly permanent and severe disempowerment of the human species?	1321	5% (IQR 19%)	16.2% (SD 23%)
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?	661	10% (IQR 29%)	19.4% (SD 26%)
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?	655	5% (IQR 19.9%)	14.4% (SD 22.2%)

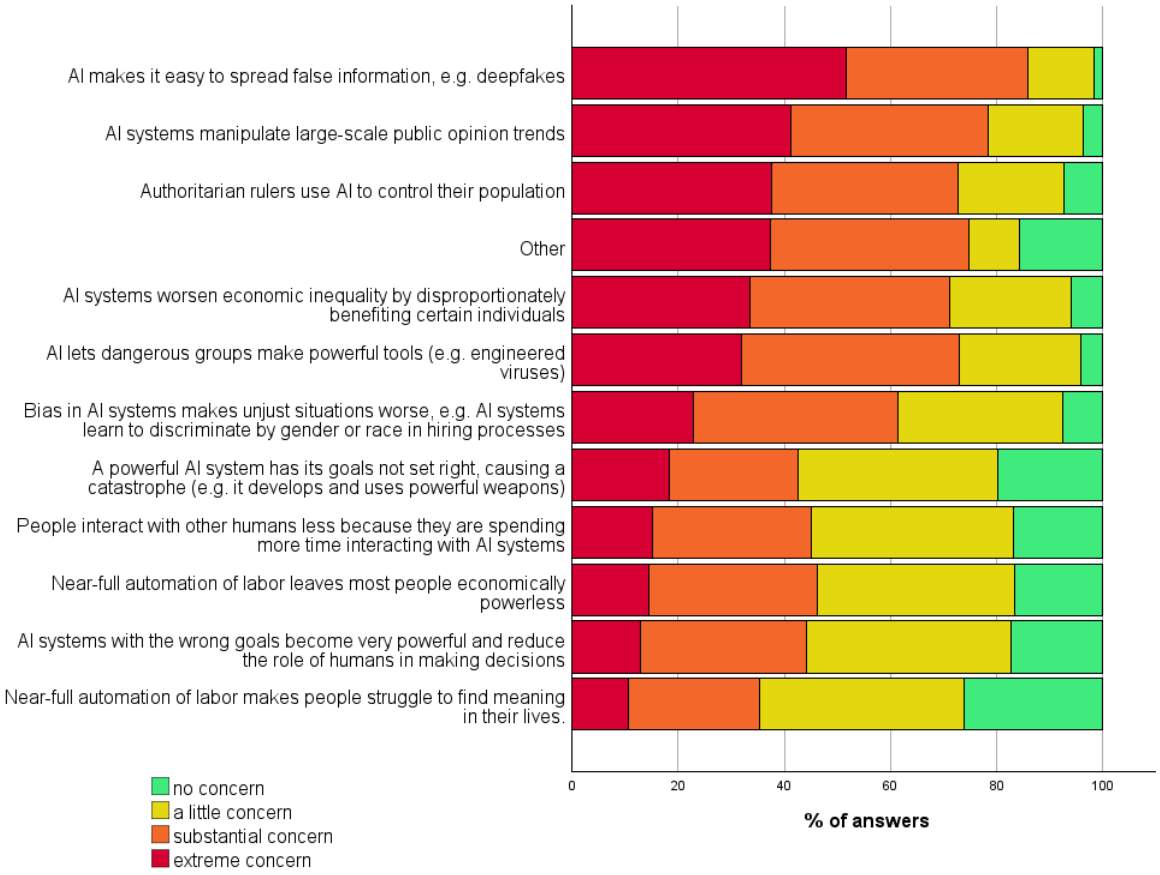
Probability of AI causing human extinction or similar	% of respondents
> 25%	10%
>33%	5%
>50%	3%
>75%	1%

Impact of High Level Machine Intelligence* on humans (N=2704)

Outcome	Median	Mean
Extremely good (e.g. rapid growth in human flourishing)	10.0%	22.6%
On balance good	25.0%	29.1%
More or less neutral	20.0%	21.4%
On balance bad	15.0%	17.9%
Extremely bad (e.g. human extinction)	5.0%	9.0%

This was defined as: “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.*”

Concern about different potential scenarios



Scenario	% of respondents with “substantial” or “extreme” concern
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%
AI systems worsening economic inequality by disproportionately benefiting certain individuals	71%

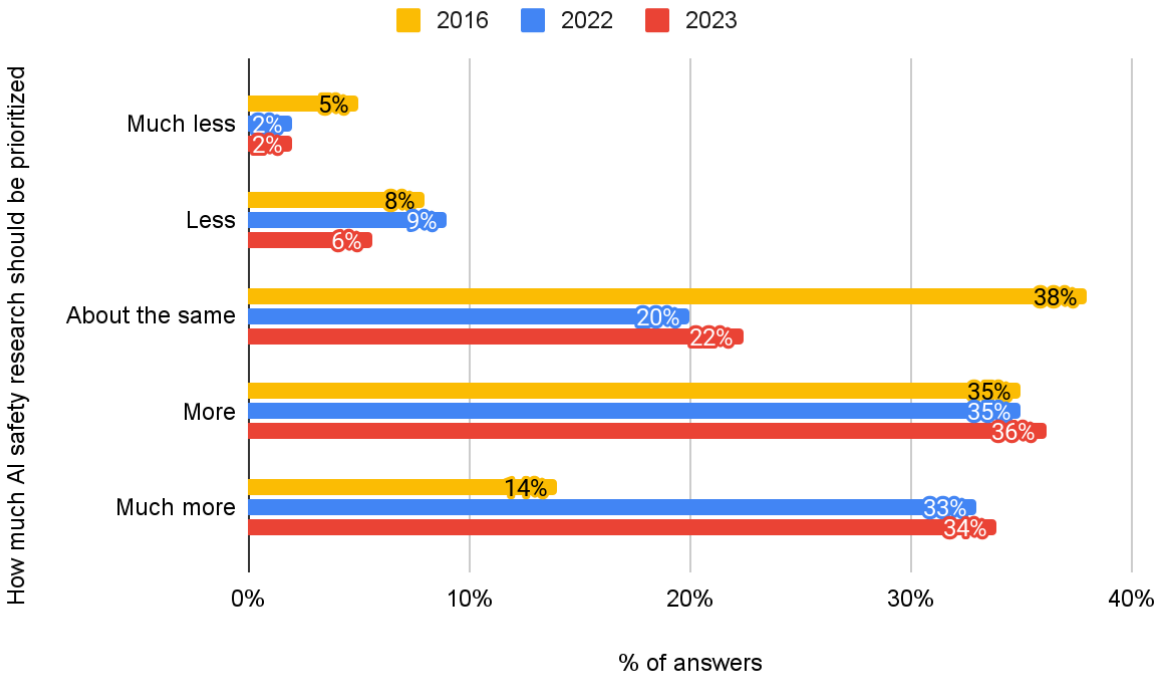
Predictions for arrival of High-Level Machine Intelligence

	2016 survey (N)	2022 survey (N)	2023 (N)
Year with a 50% chance of HLMI	2061 (259)	2060 (461)	2047 (1714)
Year with a 10% chance of HLMI	2025 (259)	2029 (461)	2027 (1714)

Predictions for AI being able to do certain tasks

	Years from 2023 until 50% chance (aggregate forecast)
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.	3.8
Write a novel or short story good enough to make it to the New York Times best-seller list.	6.8
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.	8.0

Views on AI safety



How much should society prioritize AI safety research, relative to how much it is currently prioritized? (N=1329)	% of respondents
Much more	34%
More	36%
About the same	22%
Less	6%
Much less	2%