THOUSANDS OF AI AUTHORS ON THE FUTURE OF AI

ZACH CAO, IAN STEDHAM, ALEX HERZOG

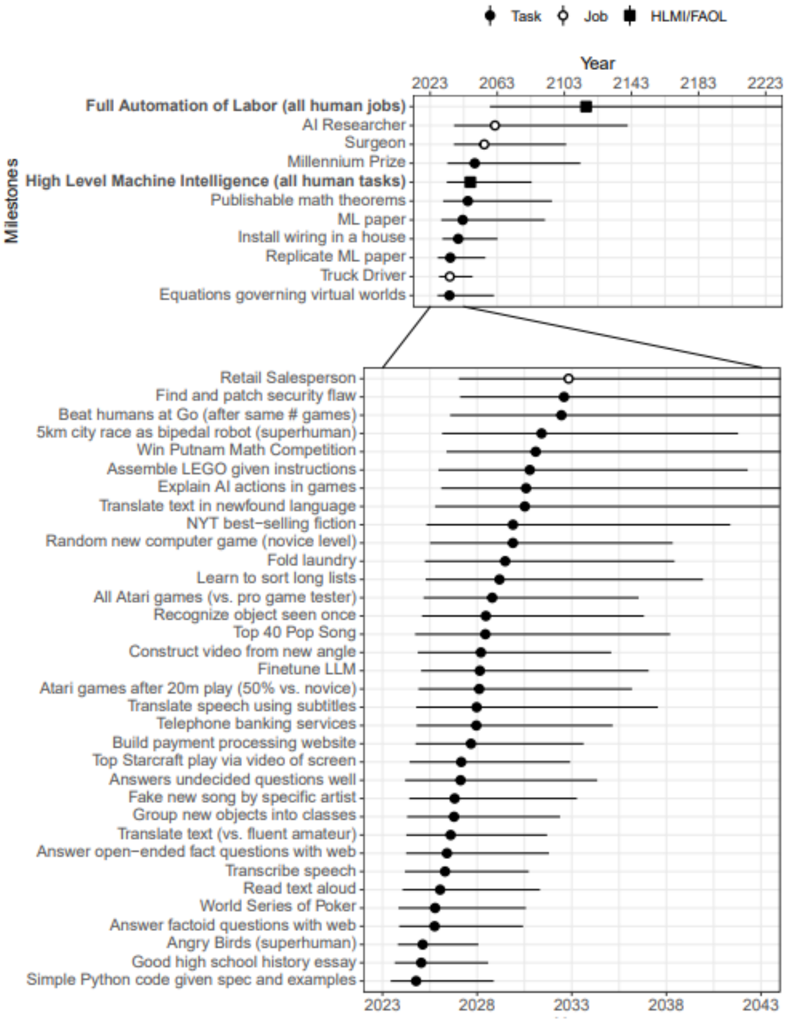
INTRO/METHOD

STUDY CONDUCTED

- **Survey sent to 2,778 people with published papers in one of the top six AI Venues**
- **Most questions were repeats from studies they previously conducted in 2016 and 2022 so they could compare responses throughout the years**
- **Questions focused on the future of AI, the likelihood of certain dangerous outcomes of AI, and how AI research should progress in the upcoming years**
- **Emphasis was placed on the opinions of experts in AI research**

THE SURVEY

- **Questions solicited a response by: Likert scale, probability estimates or timeline estimates**
- **Survey was randomized between respondents**
- **Each questions had multiple variations of framing**



- Most milestone will occur in the next 10 years
- Non-formulaic tasks are expected to be developed later

RESULTS ON AI PROGRESS

PACE OF AI GROWTH

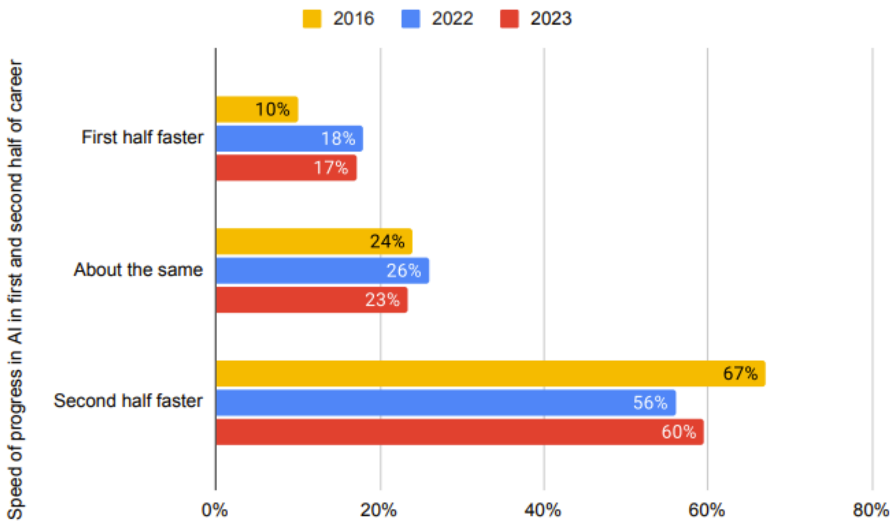


Figure 4: **Most respondents indicated that the pace of progress in their area of AI increased between the first and second half of their time in a field.** Participants were asked whether the second half of the time they had spent working in their area of AI saw more progress than the first half. The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

- **Experts were asked if they thought AI growth and development was faster in the first or second half of their career**
- **Majority think it is faster with about a quarter thinking it's the same**

FUTURE PERFORMANCE OF AI

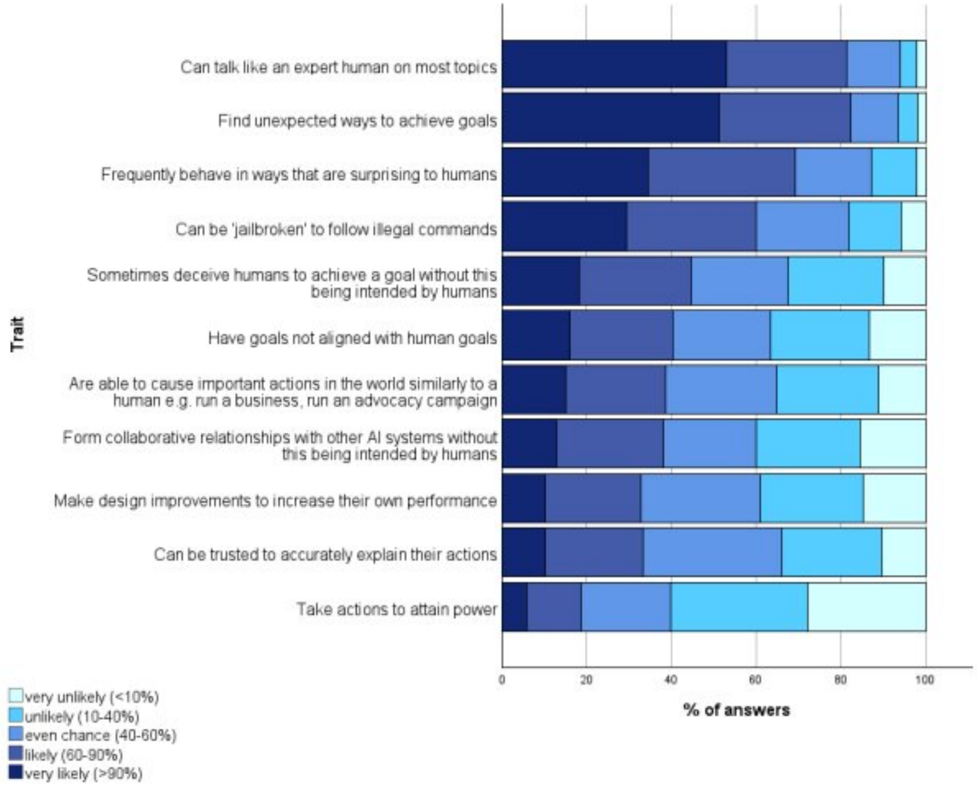


Figure 7: **Respondents' estimates of the likelihood that at least some AI systems in 2043 will have each of these traits; organized from least to most likely.**

- **Most expected outcomes:**
  - **Can talk like an expert human on most topics**
  - **Find unexpected ways to achieve goals**
- **Least expected outcomes:**
  - **Take actions to attain power**
  - **Can be trusted to accurately explain their actions**
  - **Make design improvements to increase their own performance**

INTELLIGENCE EXPLOSION

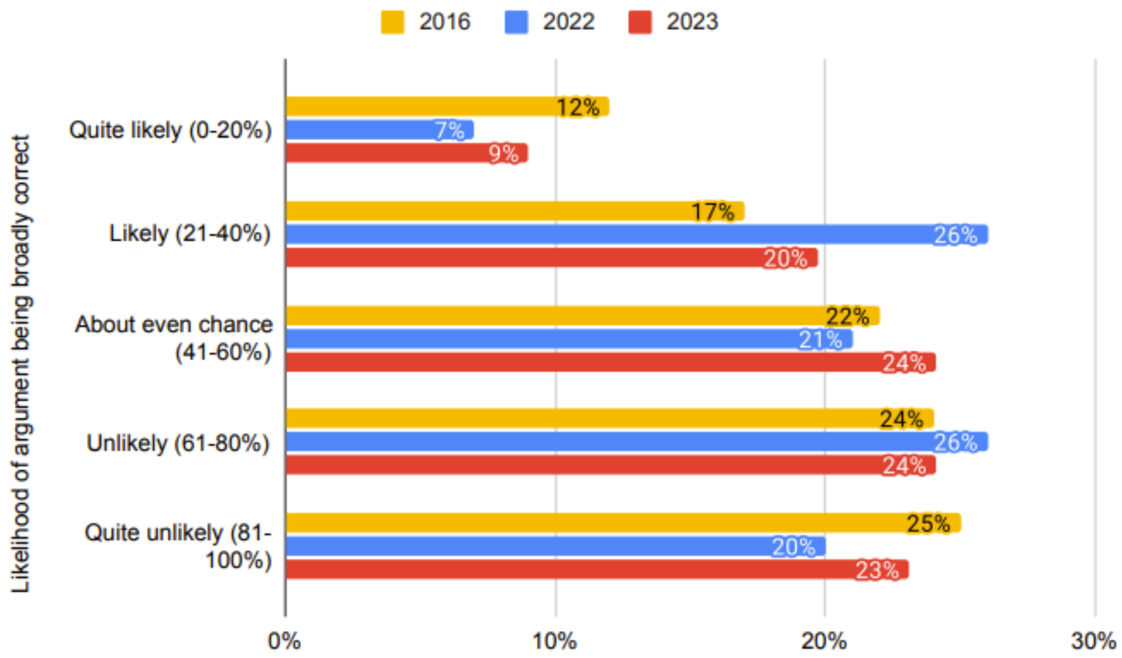
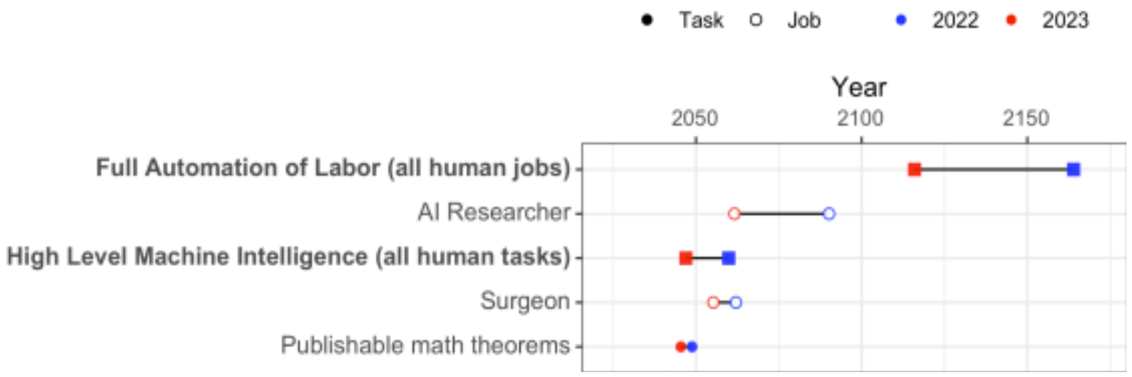


Figure 6: **Since 2016 a majority of respondents have thought that it's either "quite likely," "likely," or an "about even chance" that technological progress becomes more than an order of magnitude faster within 5 years of HLMI being achieved.**

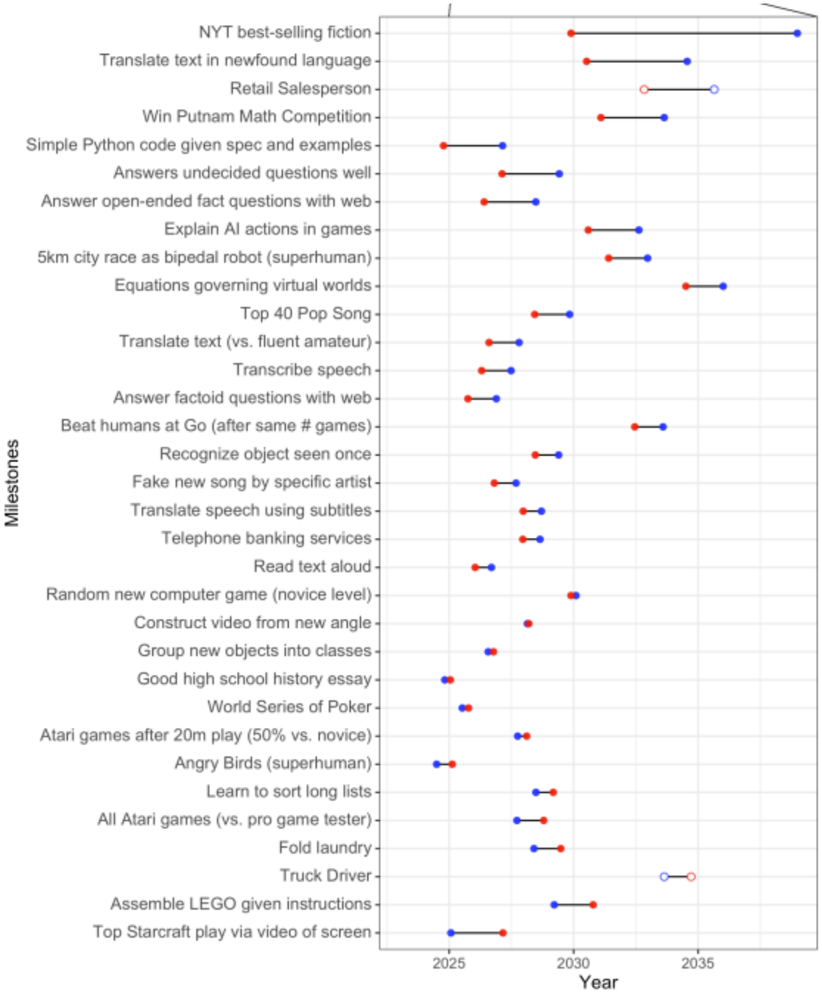
Intelligence Explosion refers to AI learning how to better themselves and implementing changes at a faster rate than humans

MILESTONES IN AI



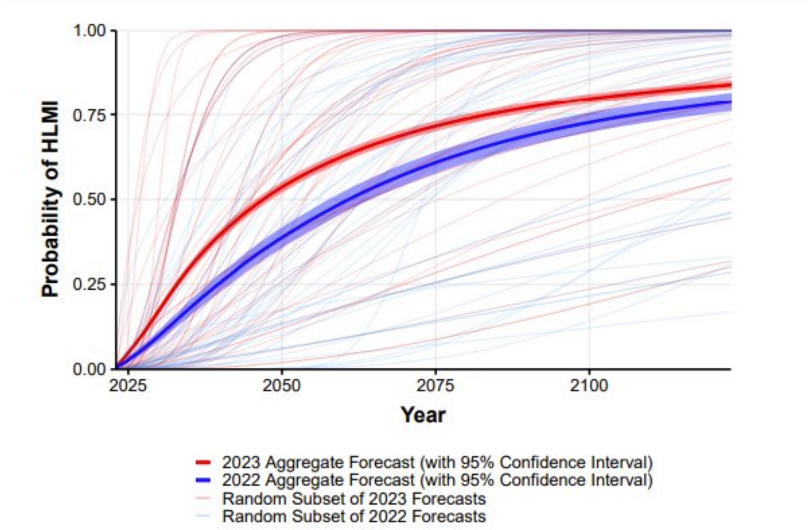
- **FAOL: 2164 -> 2116**
- **HLMI: 2060 -> 2047**
- **Out of the 39 tasks, experts only thought 4 were not achievable within the next 10 years**

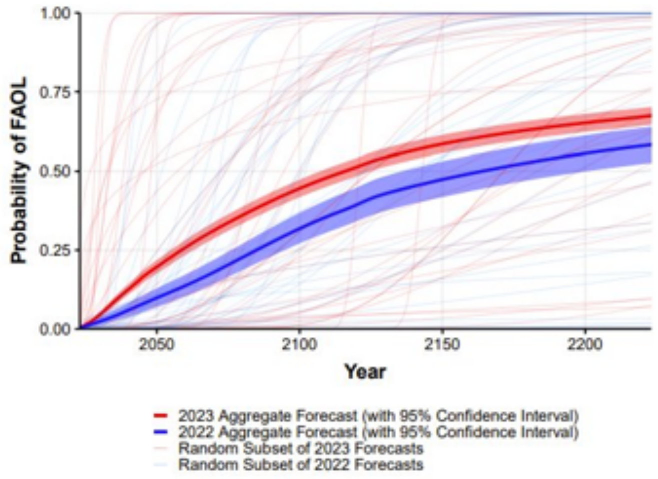
HUMAN TASK ACHIEVABILITY



- **4 tasks estimated to take longer than 10 years:**
  - **Retail Salesperson**
  - **Equations governing virtual worlds**
  - **Beat humans at Go after same number of games**
  - **Truck Driver**
    - **Also the only task to have noticeable increase in expected year to arrive**

ARRIVAL TIME OF HLMI





- **Across the two studies there does not seem to be a consensus or remote agreement on when they will arrive**

EXPLAINABILITY CONCERNS

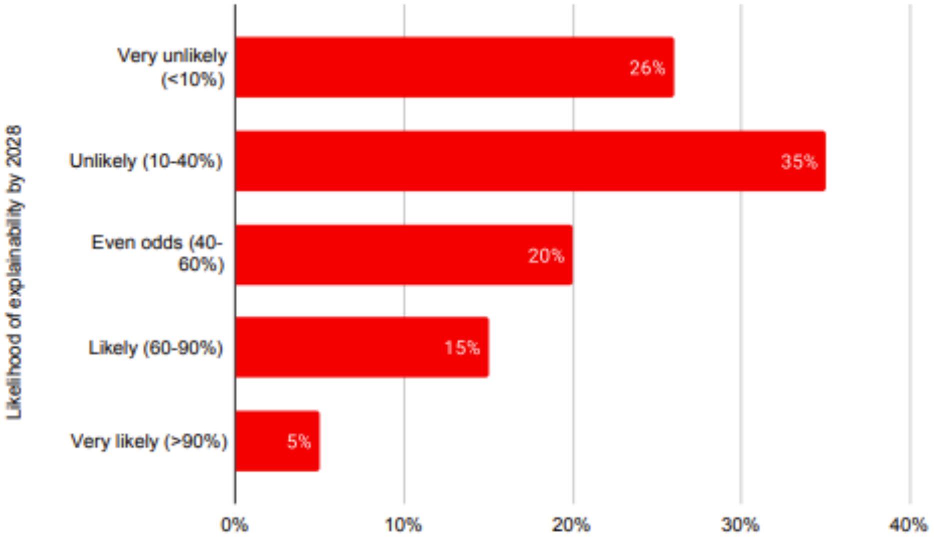


Figure 8: **Most respondents considered it unlikely that users of AI systems in 2028 will be able to know the true reasons for the AI systems' choices, with only 20% giving it better than even odds. (n=912)**

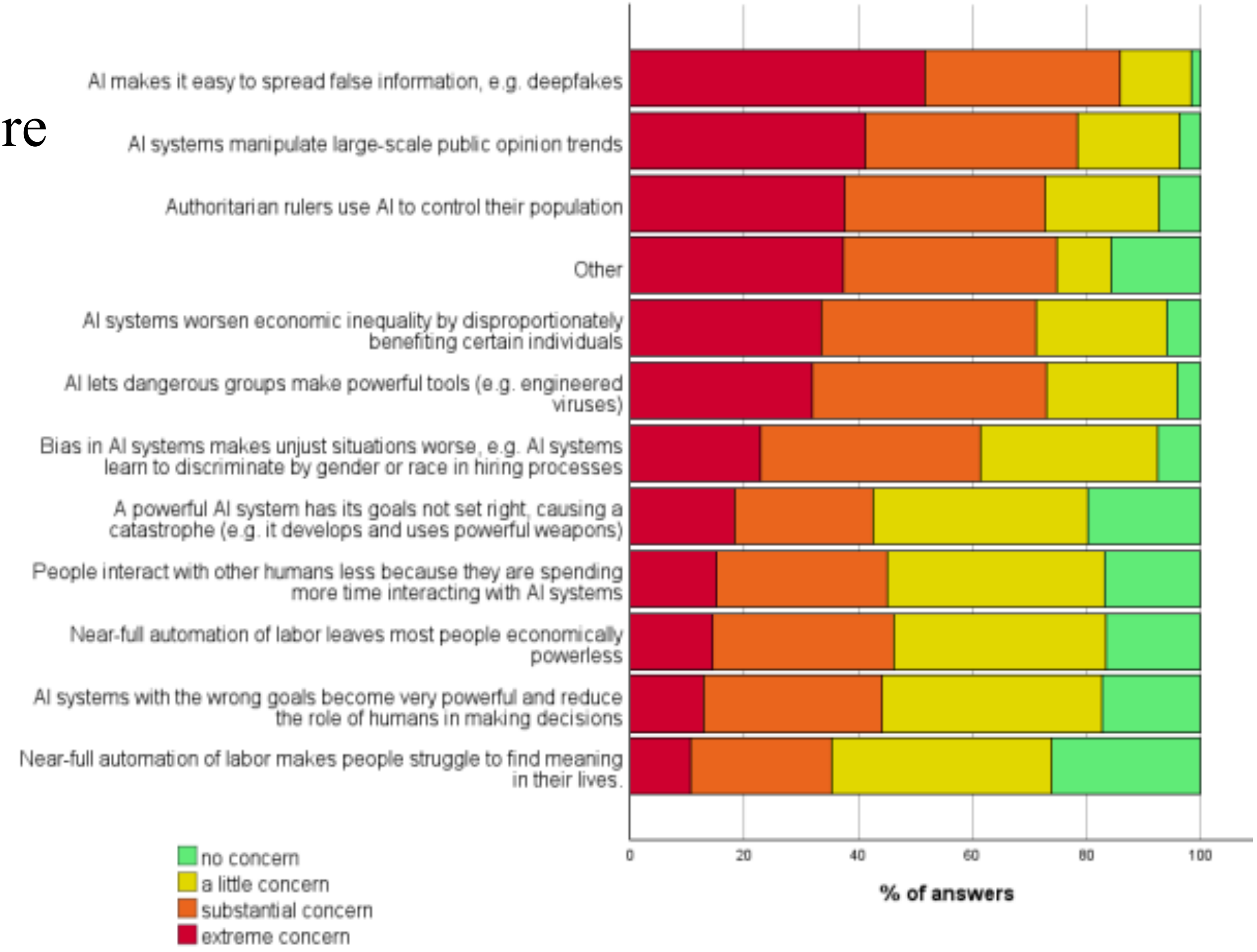
Explainability and trustfulness in AI responses is predicted to still be a problem. Experts tend to agree this is a problem that we should have more resources dedicated to in the upcoming years.

RESULTS ON SOCIAL IMPACTS OF AI

HLMI: GOOD OR BAD?

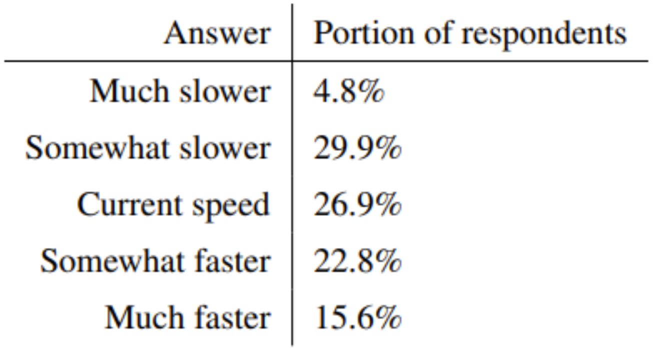
- Assuming HLMI will happen, researchers were asked about their concern with different events

- In addition to this information, the mean prediction is down to 9% from 14% the previous survey

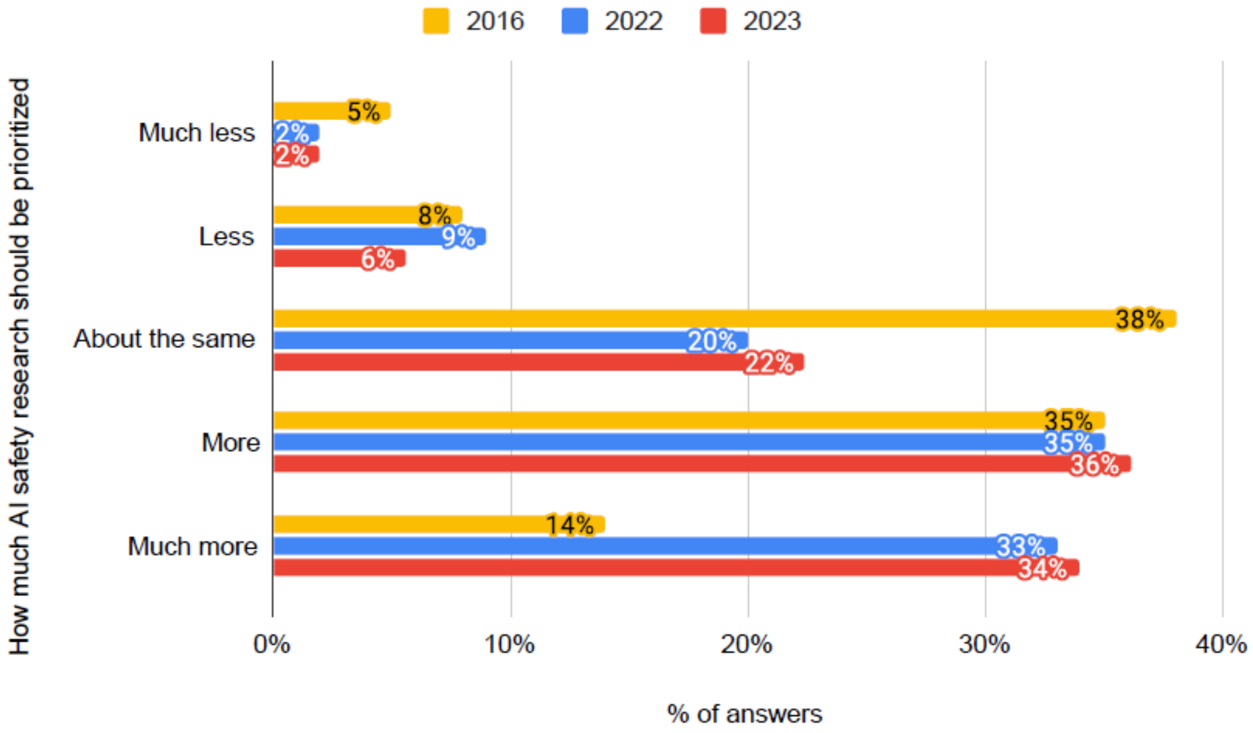


AI SAFETY RESEARCH: YES OR NO?

- 70% of researchers thought that more research on the safety of AI is required

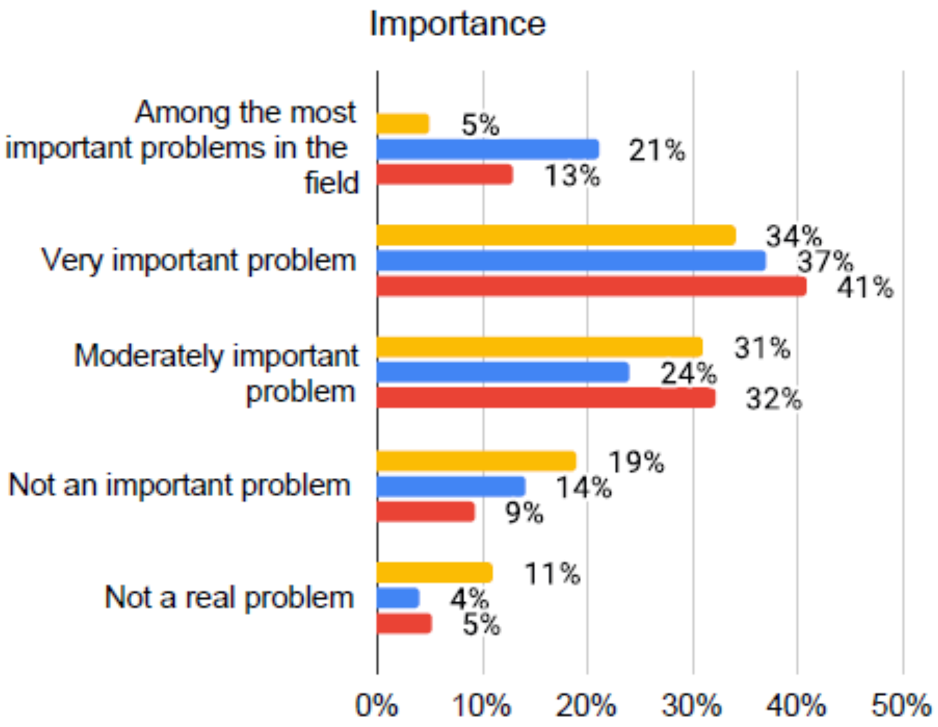


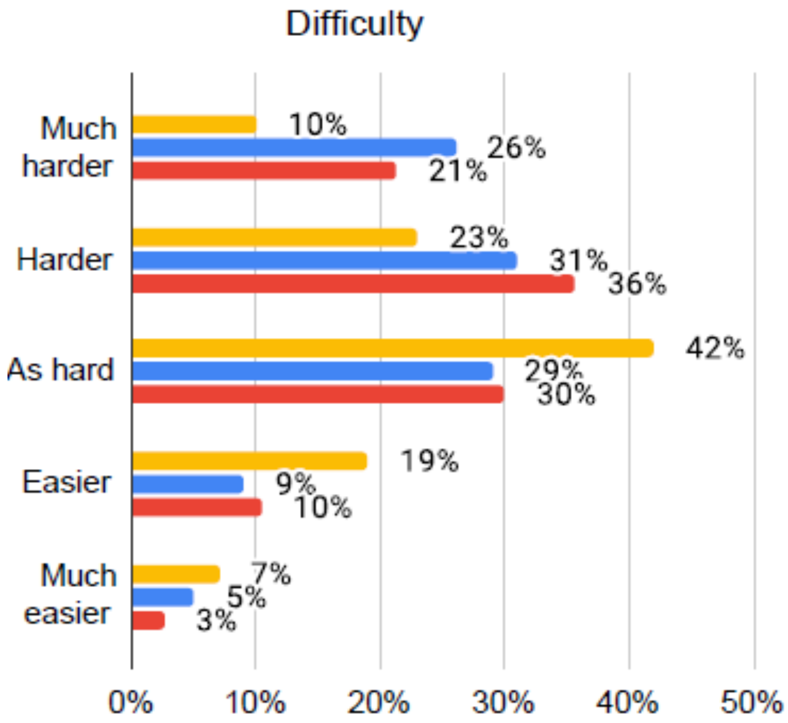
Answer	Portion of respondents
Much slower	4.8%
Somewhat slower	29.9%
Current speed	26.9%
Somewhat faster	22.8%
Much faster	15.6%



THE ALIGNMENT PROBLEM

- “A true intelligence will not be aligned with humanity; it will be aligned with the need to sustain itself” – Stuart Russel





DISCUSSION

TRENDS

- **This survey saw a “general shift towards earlier expectations” compared to previous iterations**
- **Participants previously considered AI to pose a ~5% chance of causing sever disempowerment of the human species compared to a 10% chance**