Dissecting support for a logical case on lack of safety

Published on 3 June 2023 by Nik Samoylov

I consider the logical case based on lack of safety a central avenue of persuasion for an AI research moratorium. It performed well in the narrative test and it probably needs to stand on its own in legal discourse. But how strongly do people agree with its constituent arguments?

I ran a survey of 301 respondents representative of USA general adult population between 1 and 2 June 2023. In this survey I asked the question "*On a scale of 1 to 5, how much do you disagree or agree that …?"* with a scale from *"Strongly disagree"* to *"Strongly agree"*. I also asked a series of demographic questions.

Results

The table below summarises the degree of agreement with different statements, with the margin of error (90% CI) of ±5 p.p.:

	Top 2 box (Agreement)	Middle box (Unsure)	Bottom 2 box (Disagreement)
A1: artificial intelligence (AI) poses grave dangers (such as human extinction) in the next few years	40%	27%	34%
• artificial intelligence (AI) companies can develop capabilities of AI up to human level or above	64%	26%	10%
• in a matter of a few years, artificial intelligence (AI) companies can develop capabilities of AI up to human level or above	64%	25%	12%
• *at least one of:*			
• powerful artificial intelligence (AI) can be misused for military or terrorist activities with catastrophic consequences	77%	15%	8%
• powerful artificial intelligence (AI) can go rogue with catastrophic consequences	60%	24%	16%
A2: there are no known solutions to counter extreme dangers from artificial intelligence (AI) if it is made more capable	56%	26%	18%
• you cannot simply switch off an advanced artificial intelligence (AI) system if it does not want to be switched off	49%	29%	21%
• it will be hard to contain artificial intelligence (AI) if its capabilities are at human level	66%	21%	14%
• humans do not know of a way to reliably encode complex human values in artificial intelligence (AI)	51%	32%	18%
• humans do not know of a way to reliably prevent misuse of advanced artificial intelligence (AI)	68%	19%	13%
A3: there should be a pause on advancing artificial intelligence (AI) capabilities, at least until there are solutions to make AI development safe	67%	20%	13%
A4: governments must implement a pause on advancing artificial intelligence (AI) capabilities, at least until there are solutions to make AI development safe	64%	23%	13%
A5: A moratorium is possible and can be effective.			
• an international moratorium on advancing artificial intelligence (AI) capabilities is viable	44%	35%	22%
• an international moratorium on advancing artificial intelligence (AI) capabilities can be effective	48%	31%	20%
Average	58%	25%	16%

In this hierarchy of beliefs, I hypothesised that A1 and A2 can be justified by the supporting statements.

A1

	Agreement with main statement	Lack of agreement with main statement
Agreement with all supporting statements	27%	21%
Lack of agreement with all supporting statements	13%	40%

Hit rate	**66%**

A2

	Agreement with main statement	Lack of agreement with main statement
Agreement with any supporting statement	54%	33%
Lack of agreement with any supporting statements	1%	13%

Hit rate	**67%**

The hit rates are better than chance, but not very high. This can indicate either limited validity of the hierarchy or that people "do not connect the dots".

Implications for public advocacy

- It tentatively seems that overall, more people are unsure about the different statements than disagree with them (note: this is the direction of bias introduced by questionnaire). It gives an indication that people's minds can be changed.
- The imminence of AI danger (A1) is the most disagreeable statement in this hierarchy, with a large number of people unsure about it. Convincing the public of it should be a priority, but logical arguments alone may be insufficient.
- Responses regarding A2 suggest the public are not familiar with the state of affairs in alignment research and, in particular, corrigibility. These should be explained widely in an understandable way.
- There is a good amount of support for a pause at the national level, but doubts exist for about an international moratorium.

To re-phrase, these findings suggest:

1. creating urgency around AI danger,
2. explaining uncontrollability,
3. promoting optimism for international cooperation.