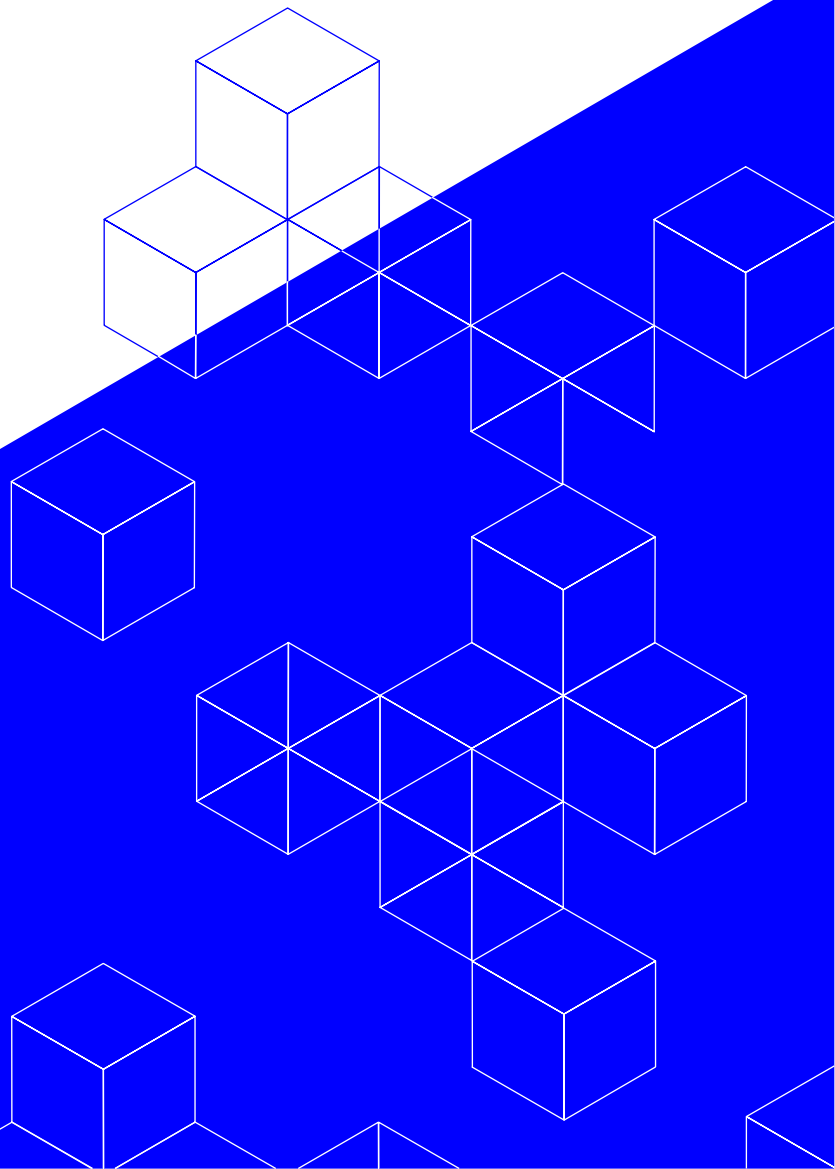


**Existential Risk
Observatory**

TRENDS IN PUBLIC ATTITUDE TOWARDS EXISTENTIAL RISK AND ARTIFICIAL INTELLIGENCE

3 April 2024

Ayushmaan Sharma



Summary.....	4
Objectives.....	4
Methodology.....	4
Trends.....	5
Base Rates.....	5
Predictors of Concern.....	6
AI Pauses and Bans.....	7
Governmental, Private, and Military Uses of AI.....	8
Survey Results.....	10
Responses to “Should the government enforce a ban on the training of AI systems more advanced than GPT-4, if the labs do not implement a quick enough pause themselves?”	10
Responses to “How concerned are you about human extinction caused by artificial intelligence?”	10
Responses to “Do you think that the development of artificial intelligence should be slowed down, accelerated, stay the same, or paused entirely?”	11
Responses to “Do you agree or disagree to the following statement: If people ever lose control over advanced AI systems, they can just turn them off.”	11
Responses to “Should AI labs halt the training of AI systems more advanced than GPT-4 for a minimum of six months?”	12
Responses to “Should smarter than human AI be banned?”	12
Responses to “Should there be a democratic vote before companies are allowed to build smarter than human AI?”	12
Responses to “Should there be an international treaty that governs smarter than human AI?”	13
Responses to “Do you agree or disagree to the following statement: smarter-than-human AI is a possibility in the next hundred years”	13
Responses to “Do you agree or disagree to the following statement: AI will be powerful enough to take over from humanity in the next hundred years”	14
Responses to “Do you agree or disagree to the following statement: smarter-than-human AI may act against humanity’s interest”	14
Limitations and extensions of this report.....	14

Summary

Leading AI academics [think](#) AI may pose an existential risk to humanity. Public awareness of these risks is important since high awareness is likely to increase both risk-reducing regulation and investment in AI Safety research. The Existential Risk Observatory has held two rounds of surveys on AI existential risk public awareness in the US and the Netherlands [before](#), namely in December '22 and April '23. This report is detailing results of a third round of surveys held recently in January '24 in these countries and the UK (n=300 per country). According to the data, public awareness of existential risk grew markedly (in the US from 7% until 15%, in the Netherlands and the UK 19%) and thereby might come close to a tipping point (25% according to [research](#)). Furthermore, the share of participants who were 'somewhat concerned', 'concerned', or 'very concerned' about human extinction caused by AI has risen from 31% to 51% in the US, a majority. Support for a government-mandated AI pause has also risen in the US, namely from 56% to 66%.

Objectives

The rapid evolution of Artificial Intelligence (AI) has sparked a global debate on its potential impacts on society, the economy, and the very fabric of human existence. However, amidst AI development, little attention has been given to public opinions or discourse on the matter. Recognising the critical need for comprehensive insights into public perceptions of AI and its associated existential risks, The Existential Risk Observatory initiated a cross-national survey to gauge sentiments, concerns, and attitudes toward AI development and its long-term implications. This involved [reiterating a previous survey](#) conducted by The Observatory to track the base rate awareness of AI existential risk, established at 7% and 12% for the US as of December 2022 and April 2023. Additionally, further datapoints indicating the public opinion of AI on important markers, including regulation, non-civilian usage, and concern, were collected. By conducting this study, The Observatory aims to inform future communication strategies on improving AI existential risk awareness to the public, policymakers, and other relevant stakeholders.

Effective communication strategy and narrative framing of existential risks is essential in mobilising public support towards risk minimisation. We deem that tracking trends in support for AI decelerationism, pauses, and bans, as well as relative concern, are of special importance. We consider that this is predicated on the belief that these markers, and potential causes for such, are especially insightful for gauging public sentiments towards AI, and policies which may regulate its usage.

Consequently, by identifying a trend in the base rate awareness of AI existential risk, as well as public opinion, The Observatory intends to explore two principal questions: is AI existential risk awareness approaching a 'tipping point'? And, if so, how should we navigate it? The *tipping point of change* phenomenon is critical to understanding how societal attitudes might shift significantly in response to increased awareness and understanding of AI risks. The concept of a tipping point, which is close to [25% according to research](#), suggests that once public awareness reaches a certain threshold, it could lead to widespread acceptance of the need for stringent regulations, ethical considerations, and proactive measures to mitigate potential threats posed by AI. This shift in public opinion could, in turn, catalyse legislative and policy changes at both national and international levels, leading to a more controlled and ethical development of AI technologies.

Methodology

A total of 900 participants were recruited to complete a Google Forms survey via Prolific, equally distributed between the UK, the US, and The Netherlands (NL). Prolific was used as the preferred choice of recruiting participants due to its ease and large population pool. The Google Forms survey comprised 25 questions for the US and NL participants and 22 questions for the UK participants. The UK study was administered first, with preliminary findings prompting the addition of 3 more questions for the remaining US and NL studies, explaining the difference in survey size between geographies. No screening filters were applied, however participants must have been at least 18 years of age to be registered on Prolific. Questions included Likert scales, open numerical risk estimation questions, to Yes/No/Maybe responses.

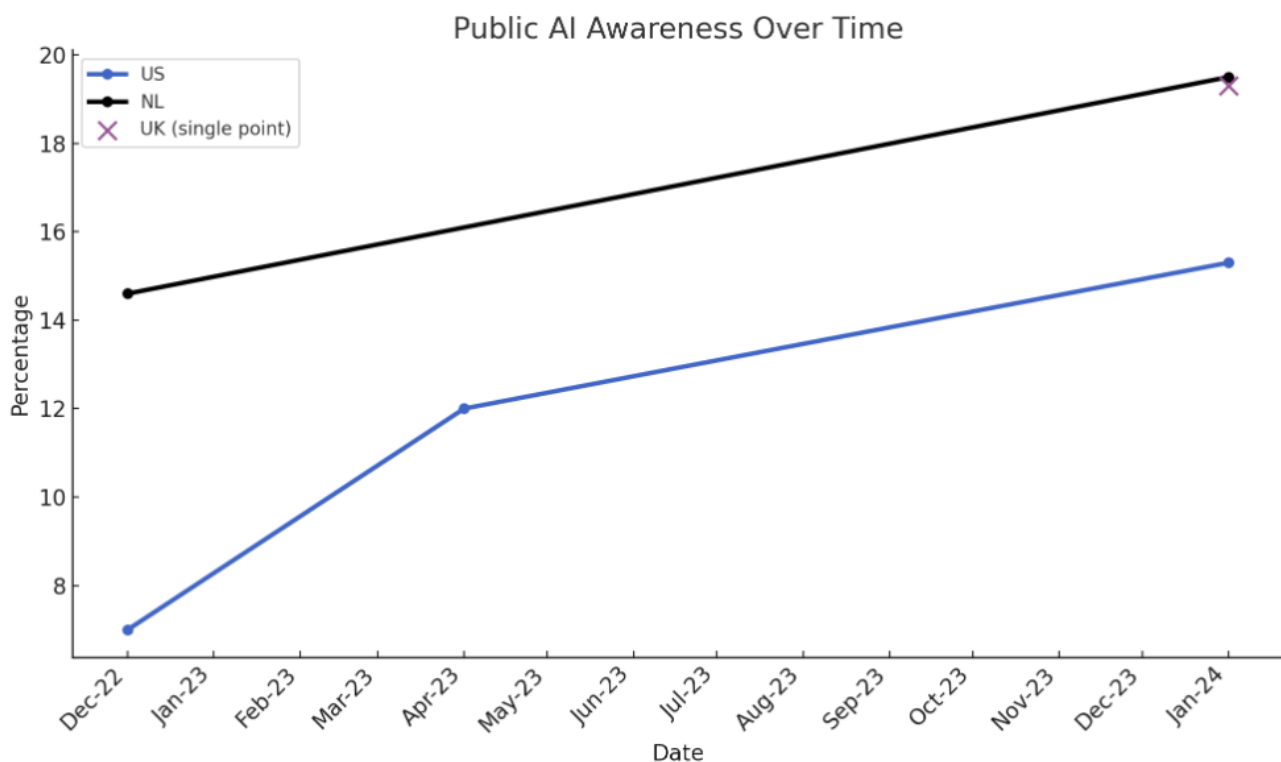
Trends

Base Rates

To calculate a base rate of awareness of AI risk within each population, participants were asked to list 3 potential causes – from most to least probable – of human extinction in the next 100 years. Given the qualitative dataset, and range of different possible responses which could indicate AI awareness, subsequent data analysis focused on reporting the proportion of responses which mentioned AI in some capacity. A base rate awareness of 19.3%, 15.3%, and 19.5% of respondents in the UK, US, and NL respectively indicated AI as one of their top 3 existential risks in the next 100 years. This is a promising increase from the previous base rates of 7% and 12% collected by The Observatory in December 2022 and April 2023, suggesting that awareness of AI and its risks is becoming more apparent.

The increase in base rate awareness of AI may be due to more widespread media coverage of existential risks in popular news, ranging from the BBC to CNBC, as previously explored by The Observatory, to key international summits such as the inaugural AI Summit at Bletchley Park, United Kingdom, to high-profile legislation such as the EU AI act. Given more current affairs, it is expected that the public will become more acutely aware of the potential usage of AI. Beyond the potential reasons for increasing awareness, those who perceive AI to be an existential risk still fall within the minority. However, the increasing base rates may suggest that the minoritarian viewpoint of AI is gaining more traction and may foreground a heightened public receptivity to mitigating measures towards AI risk.

Whether awareness of AI existential risk is approaching a tipping point cannot be decisively concluded from the data collected thus far. Assuming that societal awareness of the risks of AI is following this pattern, then efforts should shift towards ensuring the public is adequately informed of the potential harms – particularly existential threats – AI may pose. Despite this, it provides a compelling case to shift attention and communications strategies towards public opinions which assesses receptivity to AI policy and interventions.



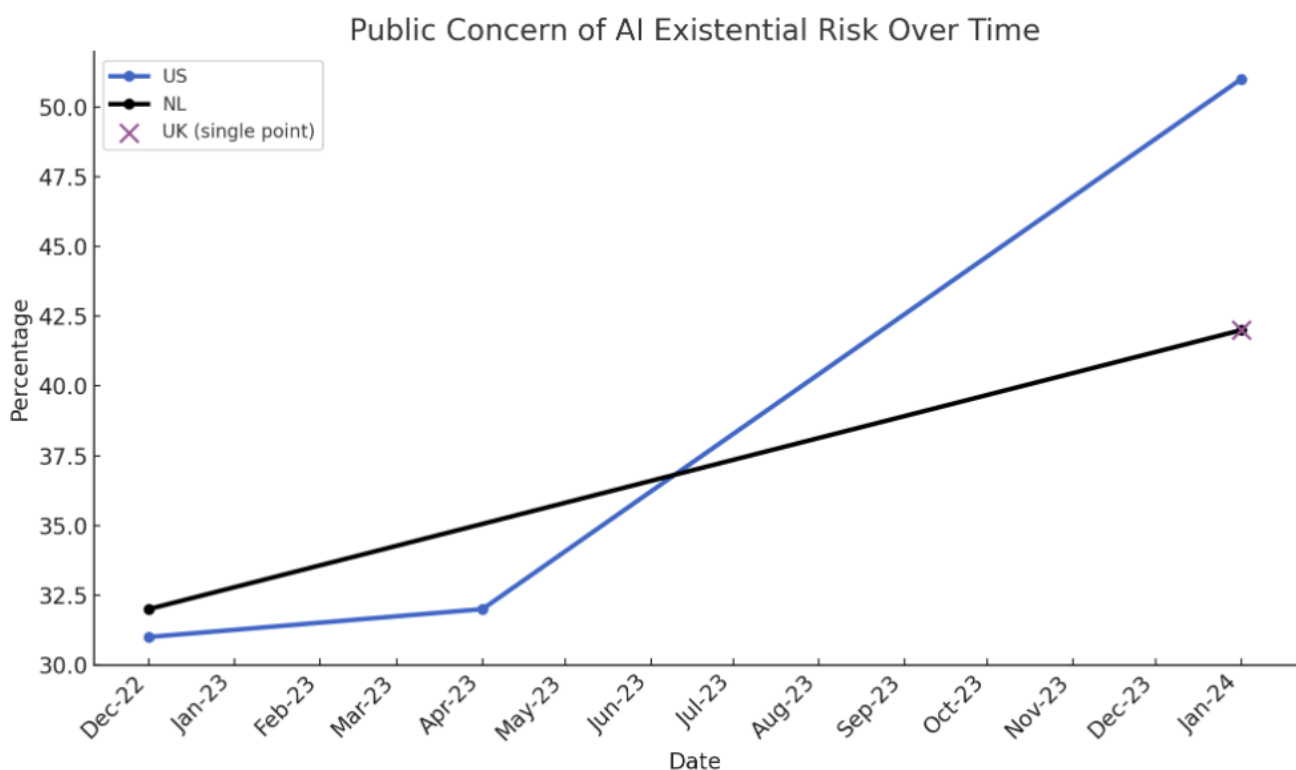
Overall trend in the percentage of AI public awareness modelled from survey data (3rd round, n=300 where n is each country).

Predictors of Concern

When considering communication strategy, it is important to determine potential causal relationships to desired outcomes. In the context of existential risk, the relative concern of the population of an emerging technology is an effective indicator of their responsiveness towards policies, or the lack of them.

A multivariable ordinal logistic regression was conducted between question responses believed to have a significant influence on participants' level of concern towards AI. The aim of this was to establish whether, for instance, those who supported a ban on superintelligent AI models also were more concerned about AI to begin with. This builds upon the work of Koen Schoenmaker, and their study "What influences concern about AI and support for slowing down its development?". We were able to reproduce their findings and confirm their "off" hypothesis – the notion that those who believed that superintelligent AI could simply be turned "off" should it pose an existential threat to humanity, exhibited little to no concern about AI risk. Targeting the "off" misconception, and the belief that the risks, beyond a certain point, are still mitigable, is an important narrative to develop to improve societal concern surrounding AI existential risk. Additionally, using the UK dataset as a paradigm, we also determined:

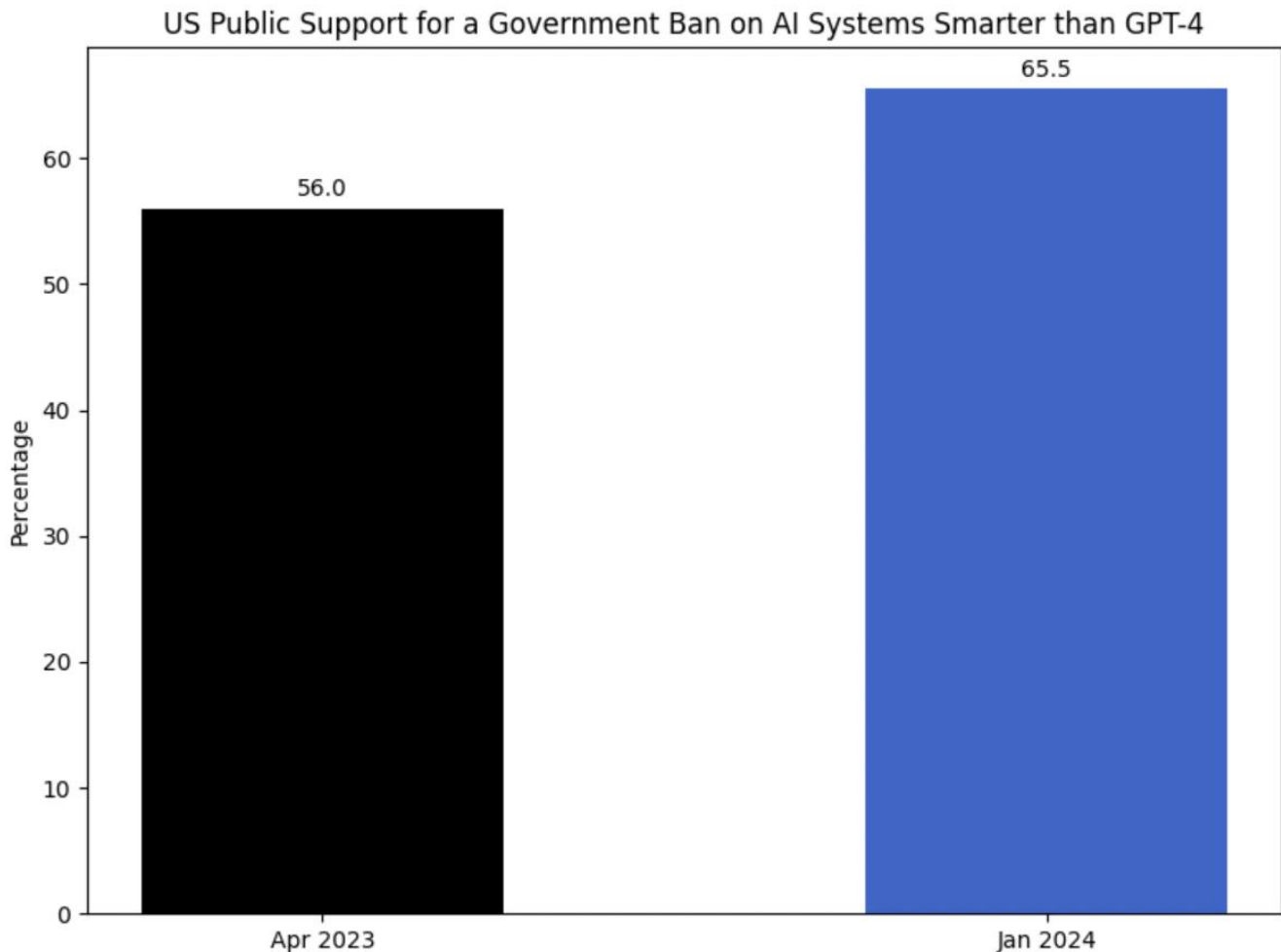
- Participants who were in favour of supporting a ban on AI indicated an elevated level of concern ($R = 0.95$)
- Participants who were in favour of slowing down the development of AI indicated elevated levels of concern ($R = 0.63$)
- Participants who were in favour of pausing AI development were more concerned about AI ($R = 0.63$)
- Prohibiting AI training on copyrighted data did not have an influence on participants' concern ($R = 0.33$)
- The belief that one could turn "off" superintelligent AI should it become an existential threat, was a strong indicator of participants' lacking concern ($R = 0.88$)
- There were no significant correlations between the demography of participants and concern

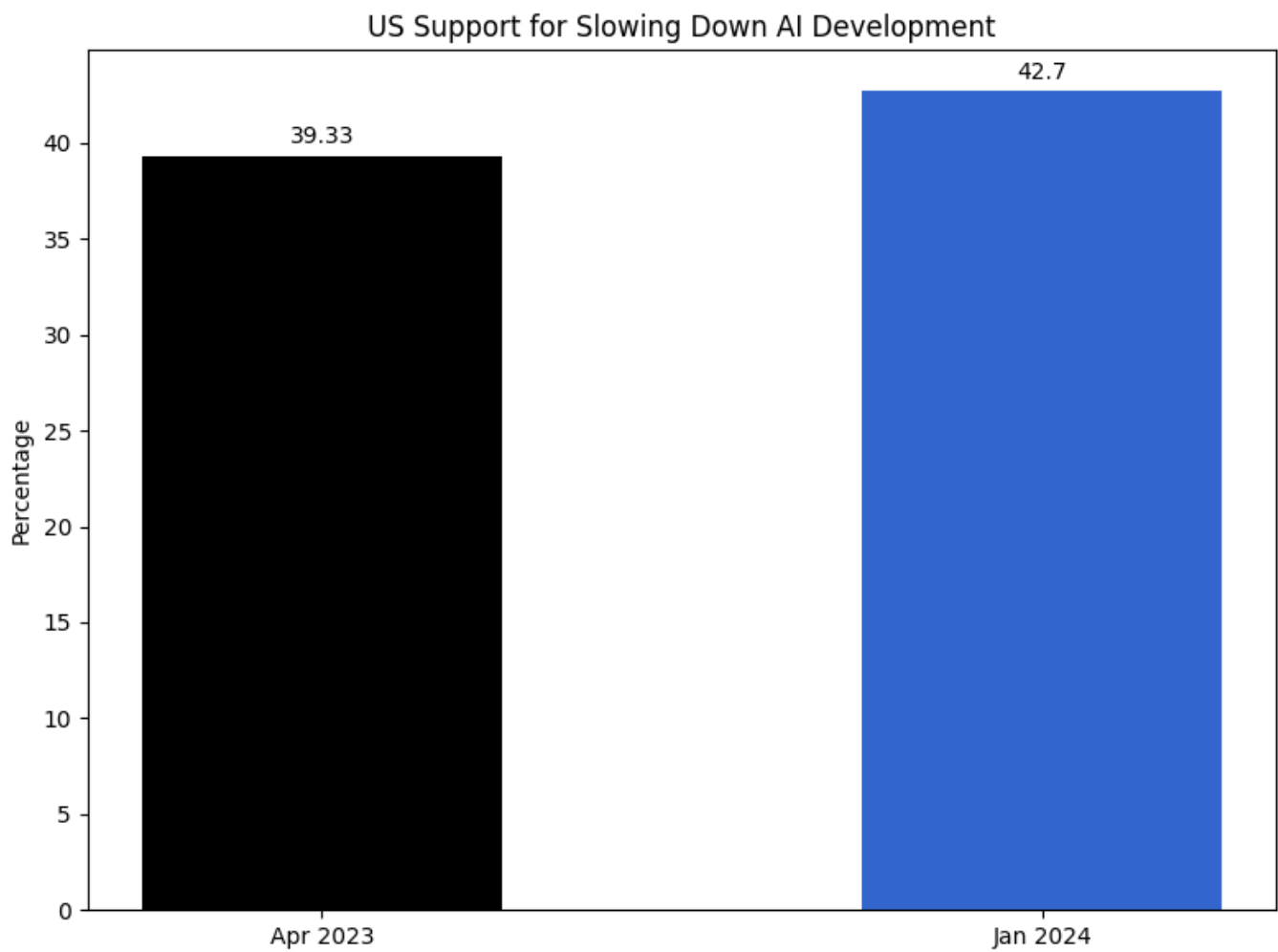


In this graph, public concern includes respondents who are 'Very Concerned', 'Somewhat Concerned', or 'Concerned' about existential risks posed by AI. For the US, concern grew from 31% in December '22 to 32% in April '23, and then increased significantly to 51% as of January '24. In the Netherlands, concern grew from 32% in December '22 until 42% in January '24. The latter was also the result for the UK.

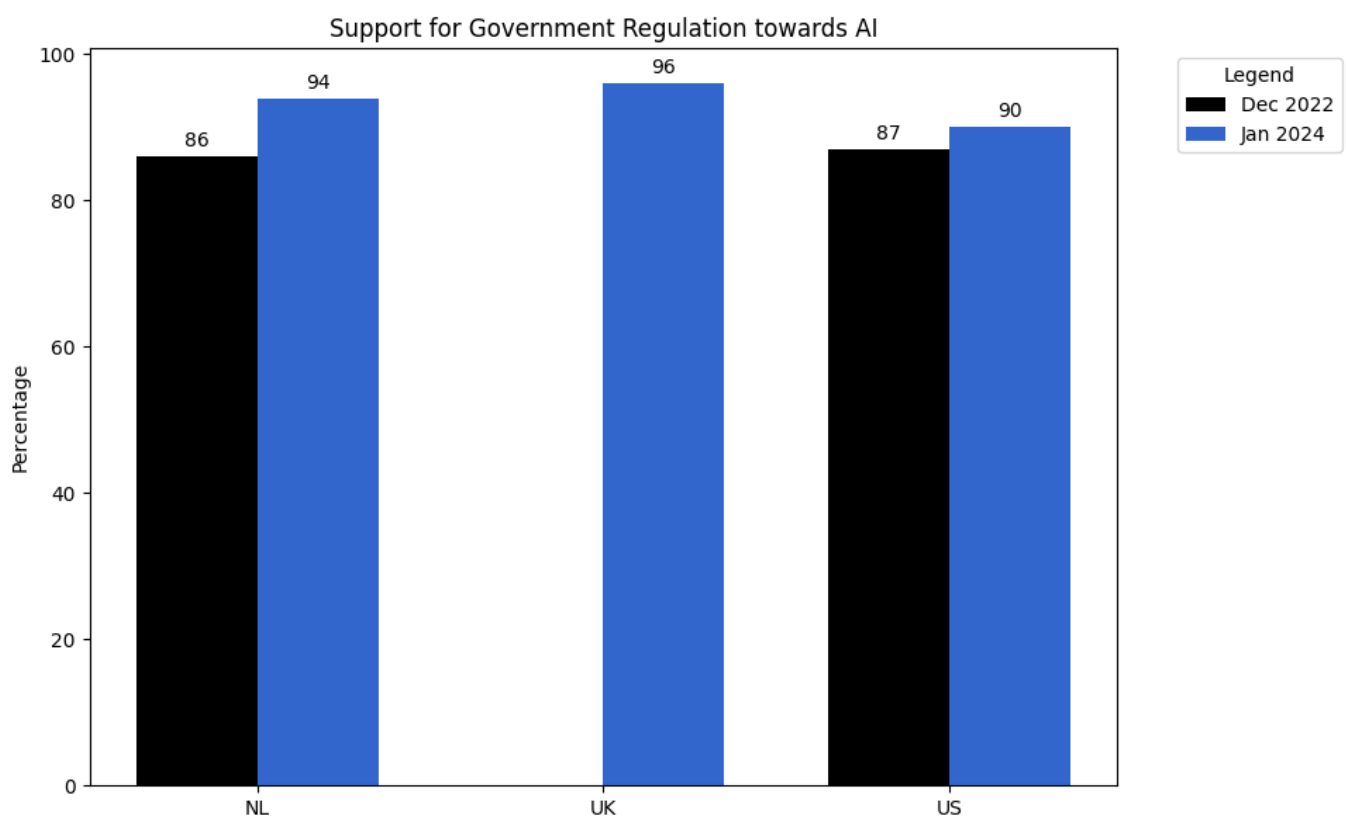
AI Pauses and Bans

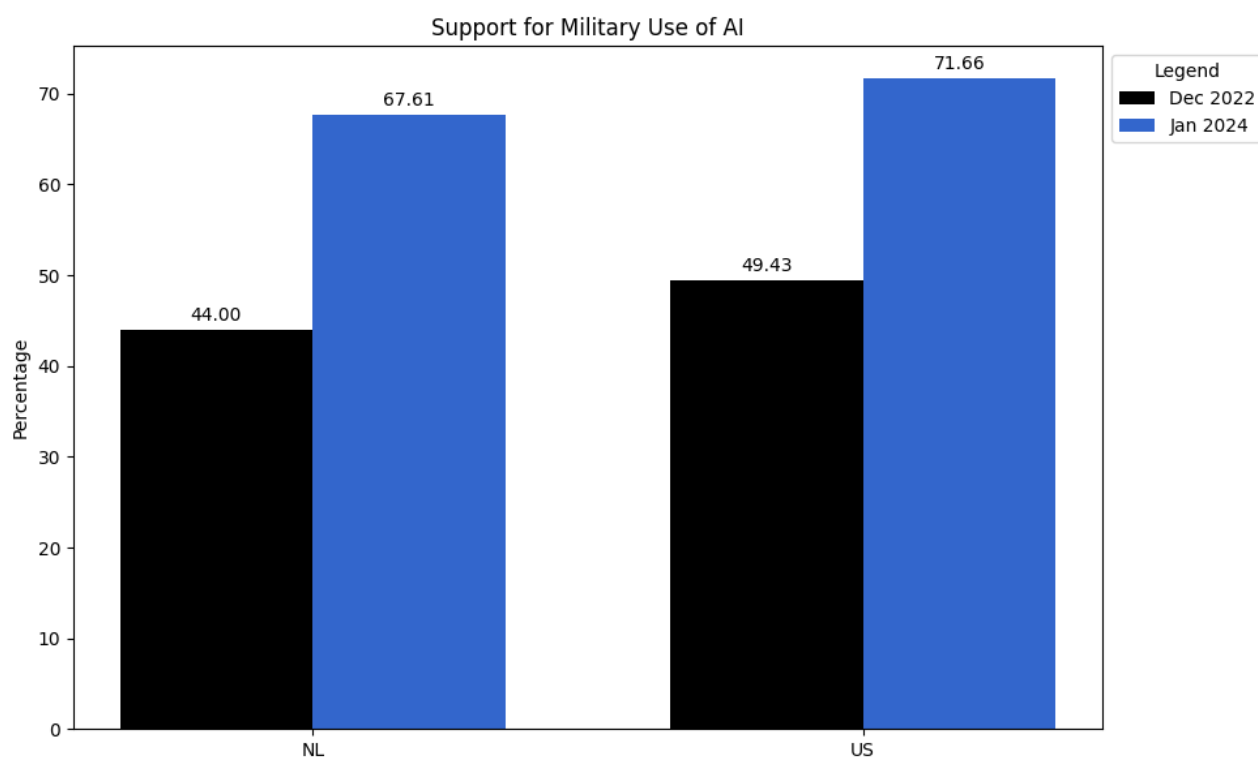
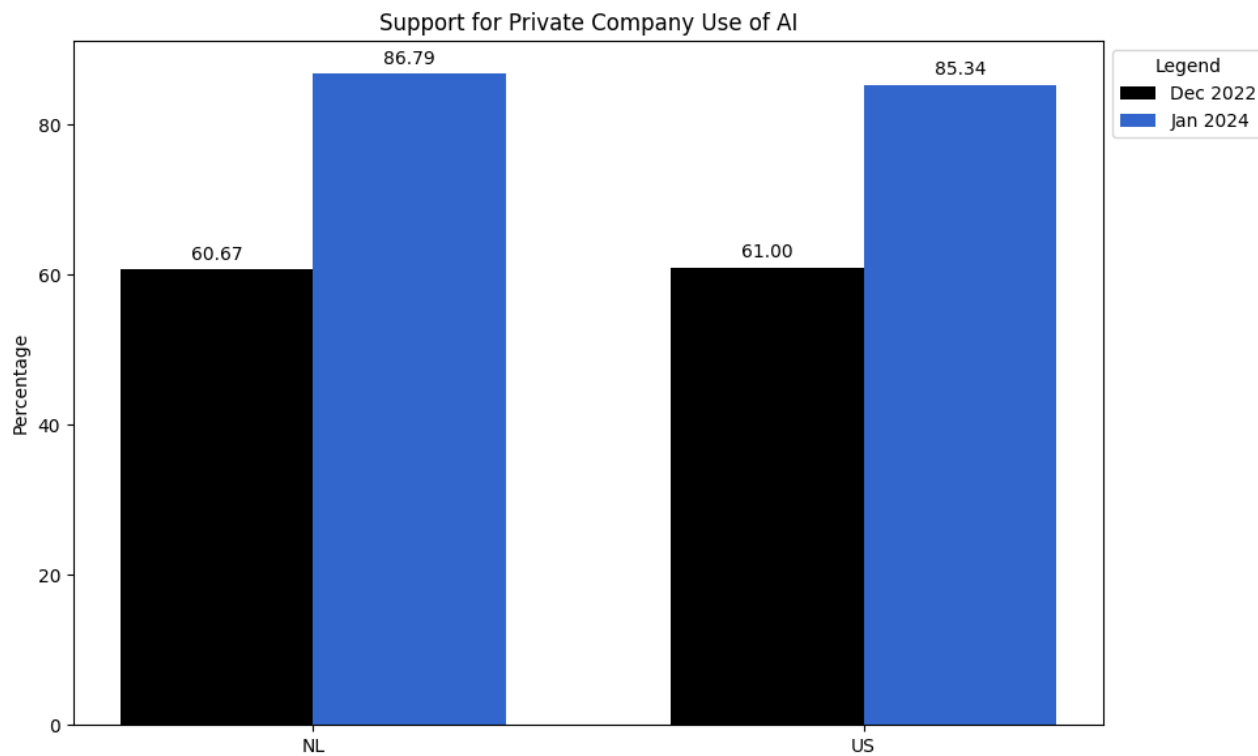
Another key finding was that in all geographies surveyed respondents who considered (responded 'Maybe' or 'Yes') banning AI and wished to see AI development stay the same or decelerate, were in the majority. This suggests that, even if the base rate of AI awareness for all countries is still in the minority, when presented with the option, the public appears to majorly favour banning smarter-than-human AI or fail to support accelerationism. This is particularly important to consider given the rise of techno-optimism, which idealises increasing AI development in spite of the risks that could be posed. It is clear that the public, removed from the context of existential risks, consider that AI development ought not to be accelerated.





Governmental, Private, and Military Uses of AI

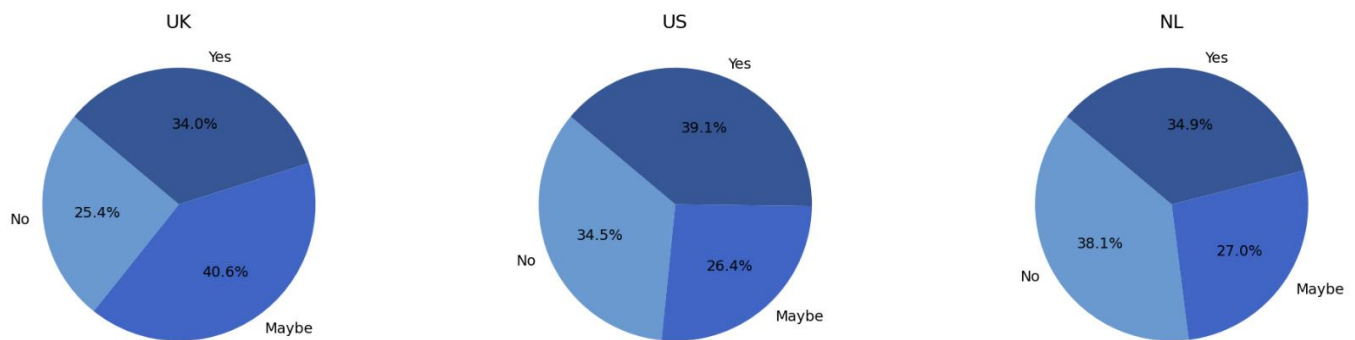




Survey Results

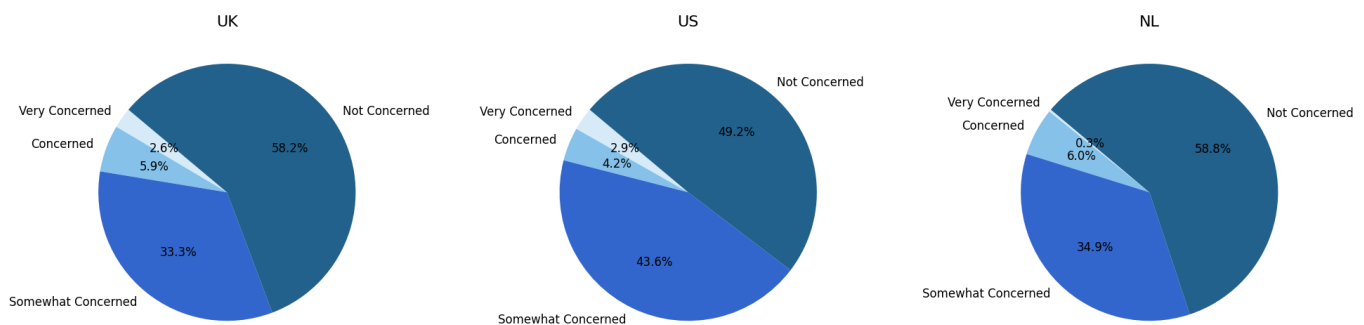
- With the recent rise of proprietary AI software being developed by private companies, the relevance of the public's overwhelming desire for AI democratisation cannot be understated. Across all geographies, respondents were in favour of a democratic vote regarding the development of AI
- There is an overwhelming consensus that, amongst all groups, when presented with the notion of smarter-than-human AI, the majority are in some agreement that it is a possibility. Although, comparatively, a smaller rate of respondents considered AI an existential risk; future strategies may benefit from focusing on the gap between a superhuman AI possible world and the nonnegligible potential for existential risk.
- Although the majority of participants responded in agreement that an AI take over is possible, to some degree, there is a lack of conviction and consensus on public estimates of the capacity of AI for a takeover. This conflicts with expert opinion in most literature, as well as open-source forecasters on websites such as Metaculus. For instance, researches from the Future of Humanity Institute, University of Oxford, estimate a [50% chance of AI outperforming humans in all tasks in the next 45 years](#) and thus having the capacity to facilitate a takeover

Ban on Training Systems More Advanced than GPT-4



Responses to "Should the government enforce a ban on the training of AI systems more advanced than GPT-4, if the labs do not implement a quick enough pause themselves?"

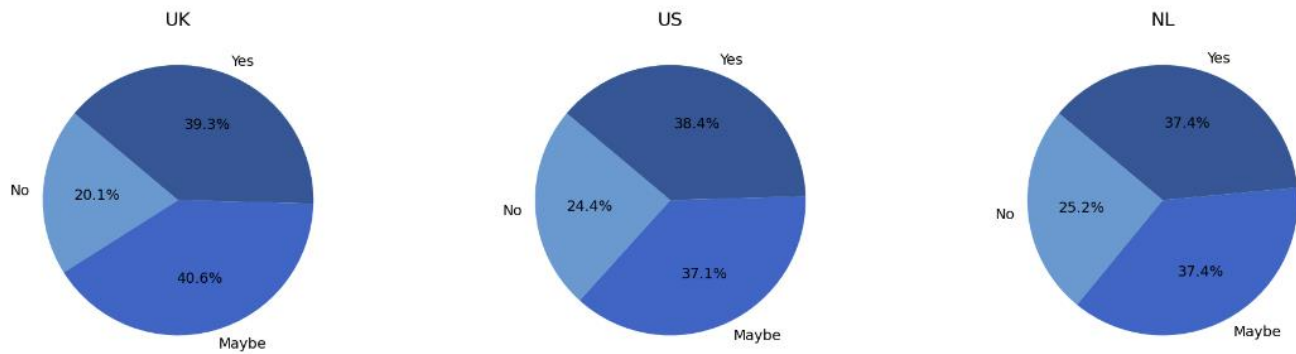
Concern for Human Extinction by AI



Responses to "How concerned are you about human extinction caused by artificial intelligence?"

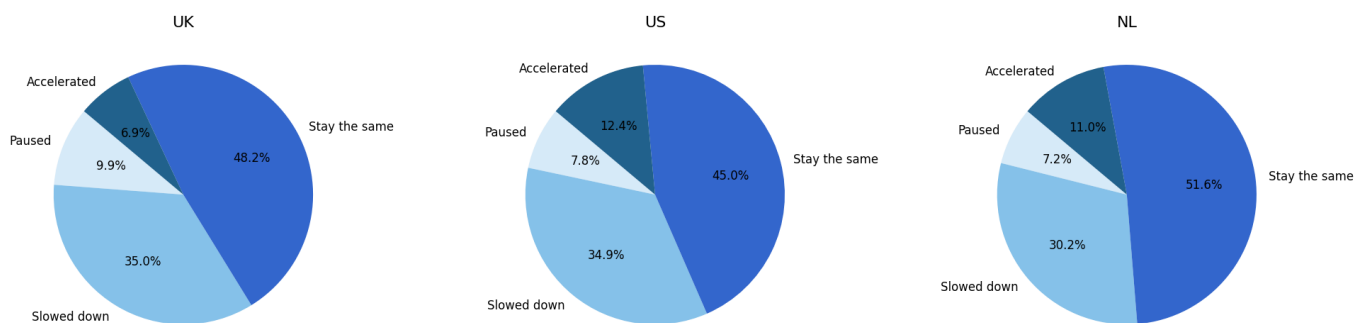


AI Training on Copyrighted Data



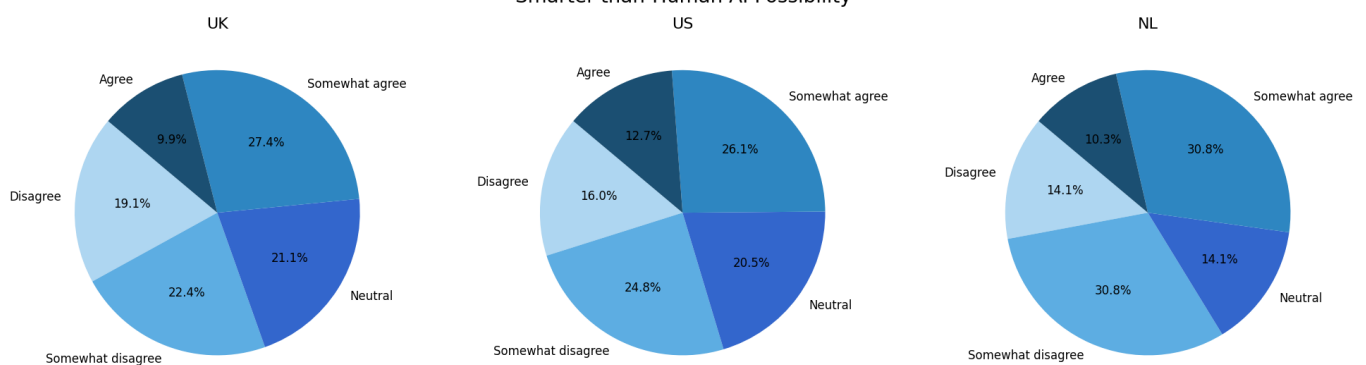
Responses to "Do you think training AI on copyrighted data should be prohibited?"

AI Development Pace Opinion



Responses to "Do you think that the development of artificial intelligence should be slowed down, accelerated, stay the same, or paused entirely?"

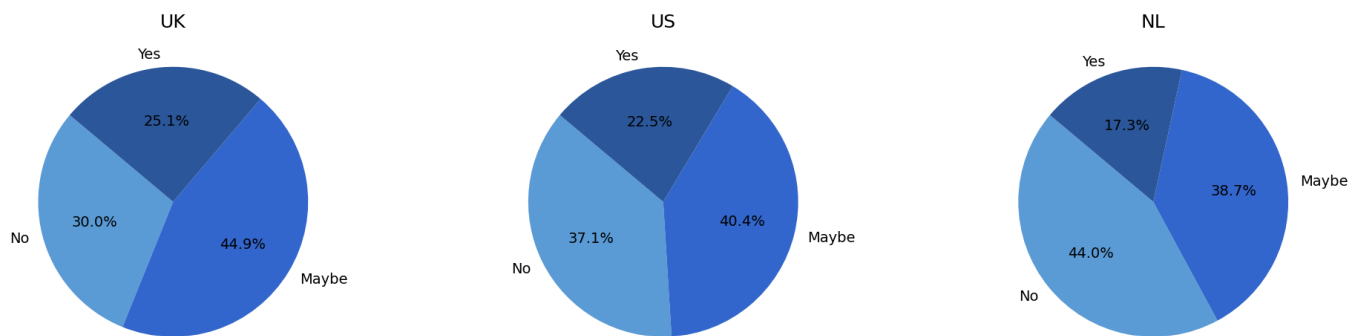
Smarter-than-Human AI Possibility



Responses to "Do you agree or disagree to the following statement: If people ever lose control over advanced AI systems, they can just turn them off."

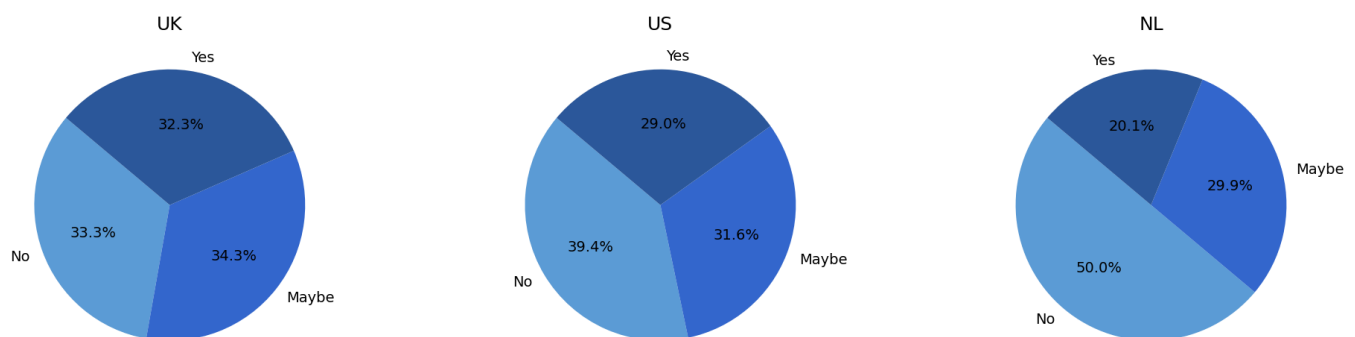


Labs Halt Training of AI Systems Smarter than GPT-4 for 6 Months



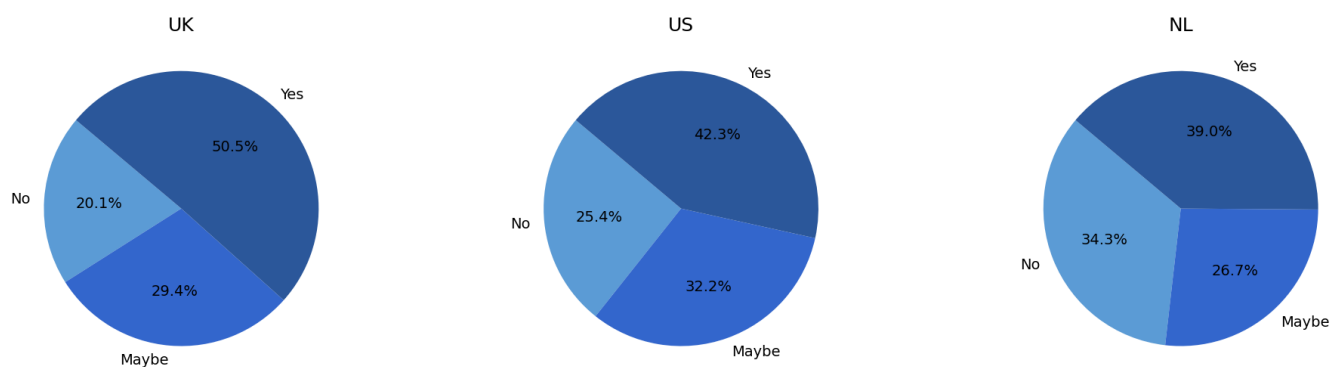
Responses to "Should AI labs halt the training of AI systems more advanced than GPT-4 for a minimum of six months?"

Ban on Smarter Than Human AI



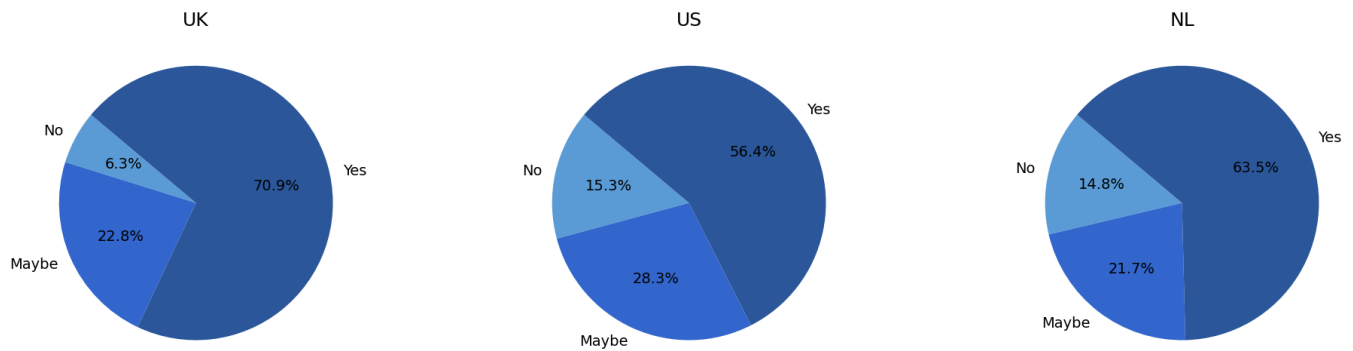
Responses to "Should smarter than human AI be banned?"

Democratic Vote on AI



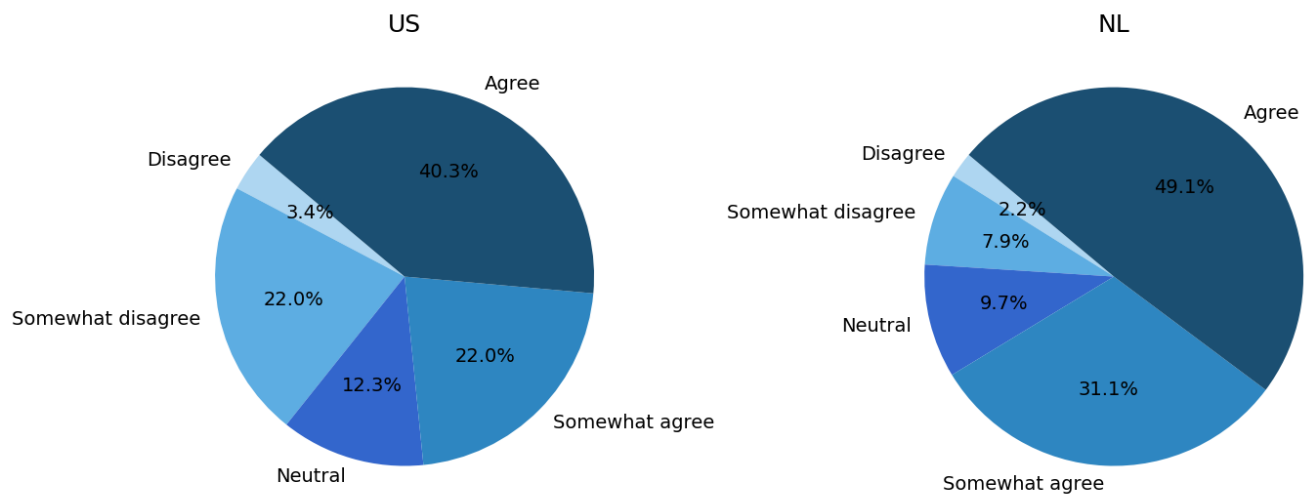
Responses to "Should there be a democratic vote before companies are allowed to build smarter than human AI?"

International Treaty on AI



Responses to "Should there be an international treaty that governs smarter than human AI?"

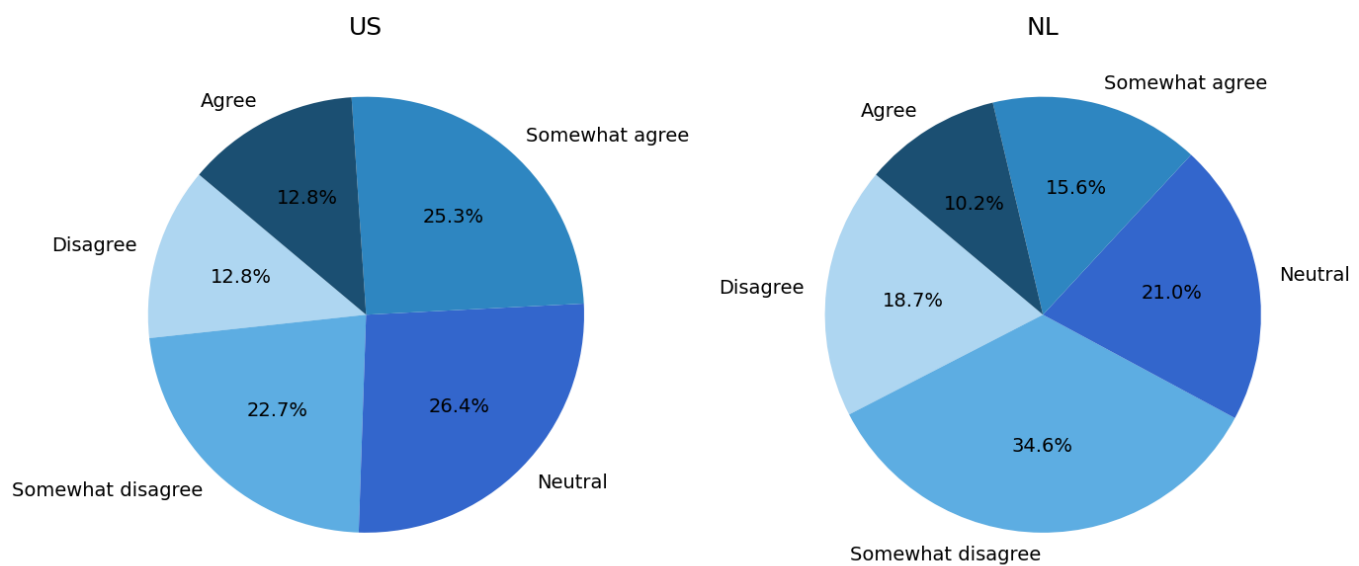
Smarter-than-Human AI Possibility



Responses to "Do you agree or disagree to the following statement: smarter-than-human AI is a possibility in the next hundred years"

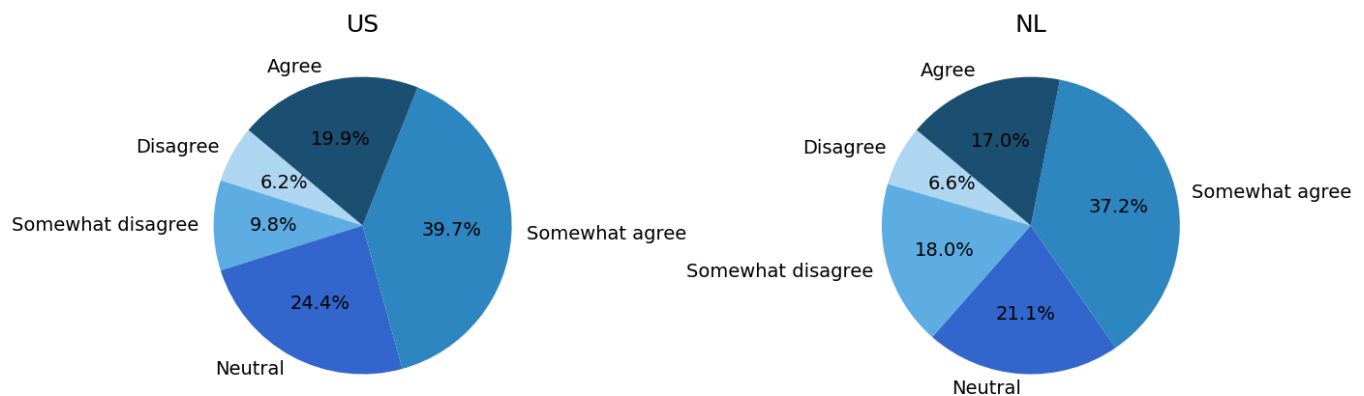


AI Take-Over Possibility



Responses to "Do you agree or disagree to the following statement: AI will be powerful enough to take over from humanity in the next hundred years"

AI Against Humanity Interest



Responses to "Do you agree or disagree to the following statement: smarter-than-human AI may act against humanity's interest"

Limitations and extensions of this report

- This survey only considered individuals from the UK, US, and NL. Further research may wish to concentrate on a particular geography, for example distribution by US or EU state
- This survey considered only 300 participants from each geography, and those who were already registered on Prolific
- This survey did not consider the relationship between common confounders towards opinion on regulation such as political skew
- This survey used categoric and subjective response markers as opposed to alternatives, such as numerical scales, which limited the extent to which regressors could be determined