The Effectiveness of AI Existential Risk Communication to the American and Dutch Public

Alexia de Roode Torres Georgiadis

March 3, 2023

Existential Risk Observatory

Abstract

The emergence of artificial intelligence (AI) has elicited both positive and negative reactions, with AI becoming a crucial component of human productivity and processes. However, there is concern that advanced AI, known as Artificial General Intelligence (AGI), may surpass human intelligence and pose catastrophic and existential risks if it is not aligned with values that benefit humanity. Despite efforts to inform the general public about the potential risks of AGI, little research has been conducted to assess the effectiveness of these efforts. This study seeks to evaluate the effectiveness of communication strategies in raising awareness of AGI risks and uses a pre-post study design that utilises surveys to measure the changes in participants' awareness of AGI risks after consuming various media interventions, namely articles or videos.

Our participants were asked to rank events they believe could cause human extinction in the next 100 years. When doing so, between 26 percent (least effective media item) and 64 percent (most effective media item) of participants newly mention AI or place it in a higher ranking after treatment. We conclude that awareness can be raised successfully using mass media for members of the general public. We also find that raising awareness is significantly more successful for female participants and participants with a bachelor degree, and we find that videos items raise awareness more successfully than newspaper articles. No significant effects were detected for age, any education level other than bachelor, country, or field of work. Our research furthermore finds that as participants become more aware of the potential dangers of AGI, the percentage supporting military work on AGI does not change significantly. This result may challenge the hypothesis that communication of AGI existential risk would constitute an information hazard via increased proclivity of the military to engage in this technology.

Keywords: Communication Effectiveness, Artificial General Intelligence, Existential Risks, Mass Media, Newspaper Articles, Videos

Contents

List of Figures

List of Tables

1 Introduction

Artificial intelligence (AI) is a technology that has spurred wavering and contrasting reactions since its conception as hopes and disillusionment have consistently accompanied its development (Crawford, 2021; Galanos, 2019). From scepticism to optimism, the range of public and expert opinion has shifted as blocks and spurs to its progress have occurred. Aside from speculation, an undeniable feature of artificial intelligence is that it has progressed immensely since its commencement. It is now embedded in many human processes and has become an essential means of productivity in the 21st century. This makes the public eye favourably placed on the capabilities and future achievements of intelligent machines (Roose, 2022). While this leads to an incentive for even faster progress in the development of artificial intelligence, there are potential harmful effects of such a trend (Roose, 2022; Russell, 2019b). As argued by computer scientist Stuart Russell and philosopher Nick Bostrom, the threat is that as computers become smarter, machines could reach and surpass human intelligence, and when that happens, if artificial intelligence is not aligned with values beneficial to humanity, it can work to our disservice - and potentially to our demise(Bostrom, 2014; Ramamoorthy & Yampolskiy, 2018; Russell, 2019a). In recent years, many tech experts have warned of the existential threats that machines can pose, such as the confinement problem in which machines expand their actions to gain access to other devices and break out of human control to pursue a goal that harms humanity (Yampolskiy, 2016). Research in artificial intelligence has been primarily focused on "narrow AI", which is aimed at great performance on specific problems, such as playing chess, rather than gen-

eral intelligence (Baum, 2017; Russell, 2019a). According to researcher Seth Baum (Baum, 2017), as of 2017 there were a total of 45 initiatives in the research and development of artificial general intelligence (AGI), of which only 4 were government-related. Even if the future is not certain, the threat from misaligned superintelligent machines is great enough that, even in small probabilities, preventing an event where artificial intelligence ends up severely harming humanity must be a priority and thoroughly researched to ensure that such a future does not happen - and if it does, that we are prepared (Bostrom, 2014).

On this note, the general public's understanding of the existential threats associated with the development of artificial intelligence is either limited or superficial. However, public awareness is critical in driving policy development and implementation, securing funding for safety work, ensuring the availability of talent to work on safety solutions, promoting general coordination, and other factors (Few et al., 2007; Khatibi et al., 2021). Thus, it is relevant to educate the general public about the potential hazards of advancing machine intelligence, and effective communication strategies are necessary for this purpose. To evaluate the effectiveness of communication interventions designed to raise awareness about the existential risks of AGI, such as articles and videos, this research aims to measure the impact of such interventions on participants' awareness of the topic. This measurement will be accomplished by conducting surveys to gather before and after data from participants on how they perceive their level of awareness after consuming the provided content.

2 Awareness of Existential Threats

Climate change is a phenomenon that few are not familiar with at this point in history (Pandve et al., 2011). Public awareness of the issue has grown significantly in recent decades, leading to its prominence in public debates and policy development and implementation. Despite this heightened awareness, action towards sustainability is progressing too slowly to ensure a seamless transition to a sustainable future, as some planetary boundaries have already been transgressed or are perilously close to being so (Steffen et al., 2015; UN Climate Change, 2015). A recent warning from the former Executive Secretary of the United Nations Framework Convention on Climate Change (UNFCCC) in 2021 underscored the urgent need for countries to meet their emissions reduction targets in the coming decades to prevent catastrophic consequences of global warming, which could result in significant instability and widespread suffering (UN News, 2021). Regrettably, existential threats are often ignored until they become imminent, and raising awareness of such threats is a lengthy and arduous process, particularly given the lengthy nature of political debates and policymaking (Ord, 2020). The global pandemic serves as an example of the consequences of neglecting hazards to human safety. The crisis emerged unexpectedly two years ago and caught the world unprepared, with inadequate awareness and rushed solutions resulting in significant loss of life (Singh, 2021; Wernli et al., 2021). Thus, it is crucial to convey the dangers of potential threats to humanity prior to their becoming urgent in order to protect it.

2.1 The Challenge of Raising Awareness

The achievements and shortcomings of campaigns aimed at raising awareness of climate change offer important insights into effective communication of imminent threats to humanity. While campaigns like Fridays For Future have been successful in reaching families, particularly younger generations who are more inclined to seek solutions, documentaries depict in detail the potential consequences of humanity's failure to act. Nevertheless, it can be challenging for many people to comprehend an event that has not yet occurred and does not have an immediate impact on their daily lives. One concept most do understand, however, is mortality. The concept of mortality is one that resonates with many individuals, and studies have shown that the idea of perceived threat is an effective tool for promoting pro-environmental behaviours (Fritsche et al., 2010; Stollberg & Jonas, 2021). Anticipating our own mortality is a unique feature of human cognition that distinguishes us from other species (Fritsche et al., 2010). However, it is important to insert existential threat awareness in an environment that supports intended behavioural change, as individuals can also resort to denial when faced with such threats (Stollberg & Jonas, 2021). The social and political context is also significant in altering established norms (Helmke & Levitsky, 2004), with gradual changes in societal values capable of driving incremental behavioural change (Helmke & Levitsky, 2004). Education is a powerful tool for driving incremental change, and in the context of climate change awareness, it remains the primary predictor of pro-environmental behaviour (Lee et al., 2015). Therefore, while the concept of mortality and the idea of imminent threat can be effective in changing behaviour, the impact of

these ideas may be limited in environments that do not support behavioural change. As a result, informing the public plays a vital role in promoting incremental changes in the socio-political context in which risk-reducing policies can be more effective.

2.2 Public Awareness of AI Existential Threats

Public understanding of the existential threats of artificial intelligence has been greatly influenced by the media and prominent tech figures that trigger alarming magazine covers (Galanos, 2019). Artificial general intelligence is a topic that generates controversial opinions: while tech giants such as Elon Musk and Bill Gates, and talented scientist Stephen Hawking state that artificial intelligence poses a threat to the future, some researchers and commentators say that such a perspective is not embodied in real technological improvements and that machine superintelligence is not a foreseeable threat (Galanos, 2019; Luckerson, 2014; Shermer, 2017). Therefore, the mainstream understanding of artificial intelligence threats has been heavily influenced by such debate (Galanos, 2019).

Nick Bostrom's book, Superintelligence, garnered worldwide attention by exposing the potential existential risk of intelligent machines, making the New York Times science best-seller in the year of its release, 2014 (Times, 2014). Meanwhile, organisations have been standing up to research these potential threats or inform the public, or both. In the field of research, the Future of Humanity Institute focuses on researching existential threats to humanity, while there are other organisations that focus exclusively on the problem of artificial intelligence. OpenAI and DeepMind are, for example, companies that, in addi-

tion to their main work as AGI capabilities companies, also focus on research aimed at ensuring that artificial general intelligence is beneficial to humanity. There are also various resources available to inform the public about the potential consequences of unregulated artificial intelligence advancements, including articles, videos, and documentaries.

2.3 Focus of this Study

Not much work has been done on measuring the effectiveness of this communication, however. One notable exception is (Warner, 2021), which concluded that existential risk messages on social media that portrayed a positive and actionable message achieved the highest engagement rate, and the topics of climate change and artificial intelligence were the most popular. Moreover, messages which related to current mainstream news seemed to trigger more user engagement. This research was limited, however, by the fact that not many users were engaging with the content in general (Warner, 2021).

This research aims to determine the most effective forms of media intervention for increasing public awareness regarding the potential dangers of artificial general intelligence (AGI). To achieve this, the study analyses ten distinct news outlets and assesses the relative effectiveness of two media formats, namely, news publications and videos. The media items effectiveness is measured through two key indicators, namely, "Human Extinction Events" and "Human Extinction Percentage," which are further discussed in the methodology section.

3 Methodology

3.1 Research Questions

This research aims to study the impact of AI existential risk communication strategies on the awareness of such threats among the American and Dutch population. The specific research questions being examined are: 1) whether AI existential risk communication promotes increased awareness, 2) how social indicators such as age, gender, education level, country and field of work affect the effectiveness of AI existential risk communication, and 3) whether the type of media intervention used (newspaper article or video item) affects the effectiveness of AI existential risk communication.

3.2 Measurement

The study design employed in this research is a pre-post design, which is utilised to evaluate the impact of artificial intelligence existential risk communication. This design involves administering the same intervention and assessment to all participants, and then measuring their responses at two points in time - one before the intervention and one after. The difference between the pre- and post-intervention measurements is used to determine the treatment effect. The research utilised the survey method for collecting data, which was administered to participants through an online Google Forms application. The survey consisted of three sections: pre-test questions, the intervention, and post-test questions.

The first section of the survey featured questions to gather demographic information such as age, gender, education, and nationality. In the second section, measurement ques-

tions were asked alongside questions about the participant's perspective on artificial intelligence. After the intervention, in the fourth section, these measurement questions were also included to compare pre- and post-intervention responses to measure the effectiveness of the treatment. Additionally, the fourth section included additional questions such as asking the participant if they felt more motivated to further research the topic after the intervention. In the third section, the intervention was presented in either of these two forms: an article or a video. The questions and interventions are listed in the Appendix in Table 3.

Each survey took around 5-15 minutes to complete and included a brief introduction about the purpose of the research and how the data would be used, as well as a consent form. Participants could only complete the survey once and no personal identifying information, such as name or email, was required. The survey included both closed and open-ended questions. Closed questions are used to assess the impact of the intervention, while open-ended questions allow participants to provide their own perspective. The effectiveness of AI existential risk communication was measured by comparing the results of quantitative questions from the pre-test and post-test sections, and the answers to the open-ended questions provide further understanding of any changes in the participant's perspective.

3.3 Operationalisation

This research aims to evaluate the effectiveness of AI existential risk communication in increasing awareness by measuring changes in participants' perceptions of the likelihood of

human extinction caused by AI. The study uses two main measurements, which are questions 5 and 6 (in Table 1 in the Appendix), to assess changes in participants' perceptions.

Question 5, or the "Human Extinction Events" indicator, asks participants to rank the events that they think could cause human extinction in the next 100 years. The research considered that there is an effect in raising awareness if participants mentioned AI after the treatment or place it in a higher ranking after the treatment. If the placement of AI remained the same before and after the treatment, or if participants did not mention AI before or after the treatment, the research considered that there was no effect in raising awareness.

Question 6, or the "Human Extinction Percentage" indicator, asks for the participants' opinion on the likelihood, in percentage, of human extinction caused by AI in the next 100 years. If there was an increase in the percentage of likelihood given by participants of human extinction caused by AI, this research considered that there was an effect in raising awareness. If there is no change or a decrease in the percentage, this research considered that there was no effect in raising awareness.

Overall, this research aims to evaluate whether or not AI existential risk communication is effective in increasing awareness of the potential risks posed by AI to human extinction, by measuring changes in participants' perceptions of the likelihood and ranking of AI as a cause of extinction before and after the treatment. The research used the "Human Extinction Events" and "Human Extinction Percentage" indicators to assess the changes in participants' perceptions.

3.4 Data Collection

The study recruited American and Dutch residents who were 18 years or older. The data was collected through the use of Prolific, a platform that finds survey participants based on the researcher's predetermined criteria. For the Dutch newspapers, participants were required to be residents of the Netherlands, and for the American newspapers, participants were required to be residents of the United States, regardless of their citizenship. Additionally, the participants were required to be fluent in English as the survey was conducted in that language. Participants were only allowed to take one of the ten surveys and had to have a minimum approval rate of 99 percent in Prolific and a minimum of 50 previous submissions in the platform. Each survey involved 50 participants, for a total of 500 participants across all surveys. The surveys were conducted in November of 2022.

3.5 Data Analysis

The data analysis for this study encompasses three main sections: (1) the key indicators "Human Extinction Events" and "Human Extinction Percentage," (2) additional indicators that explore participant perspectives on media credibility and institutional roles, and (3) the intersection between media credibility, government roles and "Human Extinction Events." The first section focuses on comparing the changes in the values of the key indicators before and after the intervention. The second section examines the variation in opinions across countries and surveys on relevant factors related to public awareness of AI existential risks. The third section investigates the influence of media credibility and government role on the

performance of the "Human Extinction Events" indicator.

To evaluate these indicators and their effects on public perception of AI existential risk, RStudio (version 2022.12.0+353) was used to create figures, tables, and models. All evaluated indicators had complete data with no missing observations (NA). Alternative responses to predetermined categories were coded for questions featuring an "other" option, but they were not explored further in the analysis, as they constituted a small portion of responses. No responses were removed from the initial sample data.

Most indicators required little manipulation of the results. For the "Human Extinction Events" indicator, a coding scheme was used to assess participants' increased awareness of AI. A variety number codes represented whether the participant placed AI before the intervention and not after, vice versa, or placed it third before and first after the intervention, etc. For the "Human Extinction Percentage" indicator, a new variable called "delta percentage" was created, which subtracted each participant's pre-percentage from their post-percentage. This variable was used as the dependent variable in the multivariate analysis conducted in Model 1. Following the analyses for each section, the discussion section presents the study's findings and their relevance to the study's purpose.

4 Results

The present section provides a summary of the data collected from the ten surveys conducted on participants, which includes demographic information such as age, gender, level of education, and field of employment. It presents a synopsis of the pre- and post-intervention

measurements, focusing on the indicators of Human Extinction Events and Human Extinction Percentage. Furthermore, the section reports the results of additional questions that were included in the survey, aimed at assessing the participants' trust in the media, their level of motivation to learn more about AI existential risks, and their perspectives and political views on the development of AI and its potential existential threats.

4.1 Description of Participants Demographics

The individuals who responded to surveys from Dutch newspapers AD and Trouw were all from the Netherlands, while those who took part in surveys from American newspapers were from the United States. It should be noted that being a citizen of either country was not a requirement to participate. On average, the Dutch surveys had a narrower age range among participants compared to the American surveys. However, all ten surveys had a majority of participants between the ages of 20 and 40, and a limited number of participants over the age of 50. Figure A in the Appendix shows the age range of participants for each survey.

In the surveys, participants were evenly split between men and women, with each group making up about 50 percent of the sample. The majority of participants had either a bachelor's degree or a high school education, and few held a higher degree such as a master's or doctorate. The field of science, technology, engineering, and mathematics (STEM) (not related to artificial intelligence) was the most commonly represented among participants in most surveys. In each survey, at least 30 percent of participants selected the "other" option for their field of work, while other fields had values ranging from 0-20 percent,

including options such as arts and media and healthcare. The field of artificial intelligence and military work were not well represented among the participants, comprising less than 5 percent of the sample. Additional information can be found in Table 4 in the Appendix.

4.2 Summary Statistics of Pre-Post Measurements

4.2.1 Human Extinction Events

The majority of participants, 92 percent in both countries, did not include AI in their initial response to the Human Extinction Events indicator. In the United States, 93 percent of participants did not include AI in their initial response, compared to 89 percent in the Netherlands. However, the results of pre and post-treatment measurements show that 40 percent of participants increased their awareness of AI, either by including it in their post-treatment response or by increasing their ranking in the post-measurement. The United States had 42 percent of increased awareness while the Netherlands had 37 percent. Figure 1 illustrates the percentage of increased awareness among different surveys, with the CNN survey performing the best with 64 percent increased awareness, followed by the PewDiePie and the first Trouw survey, both at 44 percent. The Washington Post survey had the lowest increase in awareness at 26 percent. The CNN, PewDiePie and Elon Musk surveys all used video format treatments which may suggest that videos are more effective than articles. Another possibility is that media items featuring well-known individuals may perform better as these three interventions had a public figure delivering the information. Reasons why CNN's item performed so well may include a video format intervention,

featuring the well-known figure of Stephen Hawking, or a convincing presentation by James Barrat.

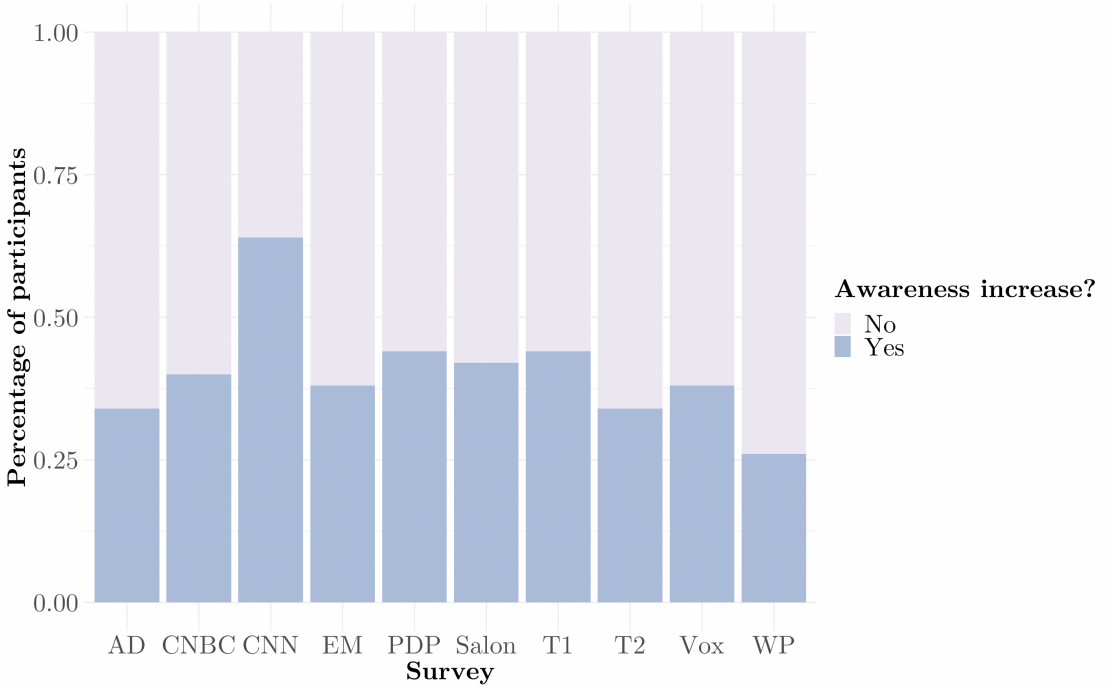

Figure 1: Percentage of participants who exhibited higher awareness after the intervention across surveys

The data in figure 2 illustrates that there is a difference in awareness levels between men and women in both the United States and The Netherlands, with the discrepancy being more pronounced in the Netherlands. The survey results show that nearly 50 percent of women from both countries were convinced, while men were closer to 25-30 percent. It suggests that there is a gap in the increase of awareness levels between men and women, with women being more likely to change their perceptions towards AI.

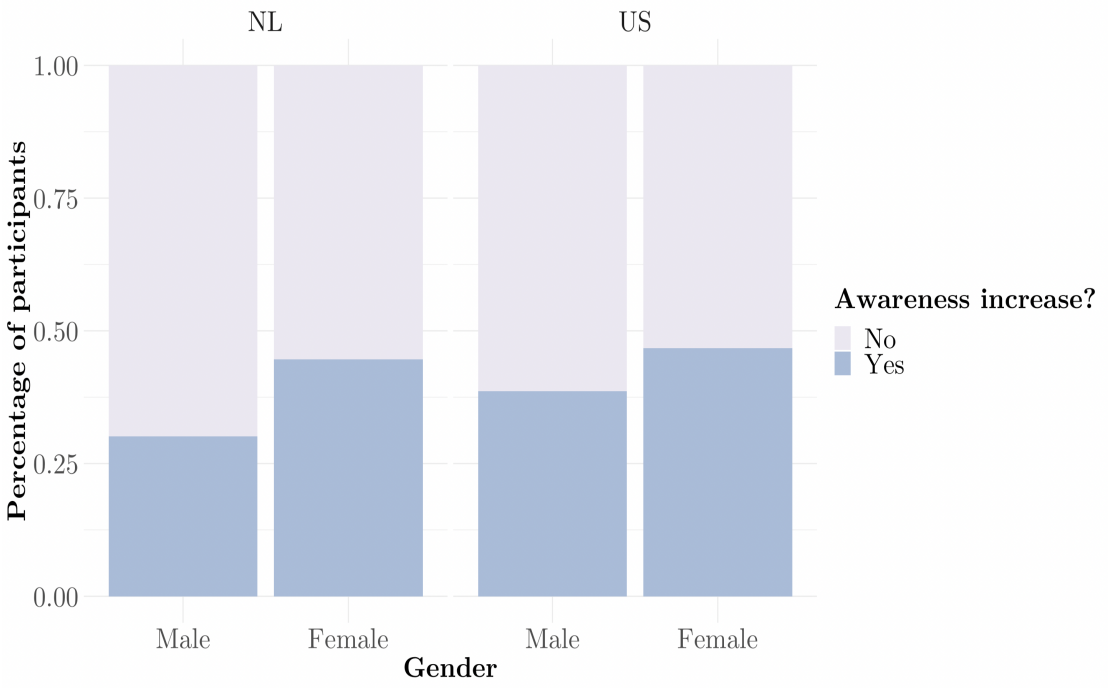

Figure 2: Percentage of participants who exhibited higher awareness after the intervention between genders and across countries

Figure 3 illustrates that participants with a bachelor's level of education had the highest increase in awareness, at 45 percent, followed by those with vocational education, at 43 percent. The other educational levels had an increase in awareness around 35 percent, with high school having the lowest increase in awareness. This may suggest that higher levels of education are associated with a higher increase in awareness. It is worth noting that while Masters and Doctorate had only slightly higher percentage than high school, it must be taken into account that there were less data for these educational levels. Therefore, it is possible that more data is needed to determine any correlation between higher education and increase in awareness of potential existential risks of AI.

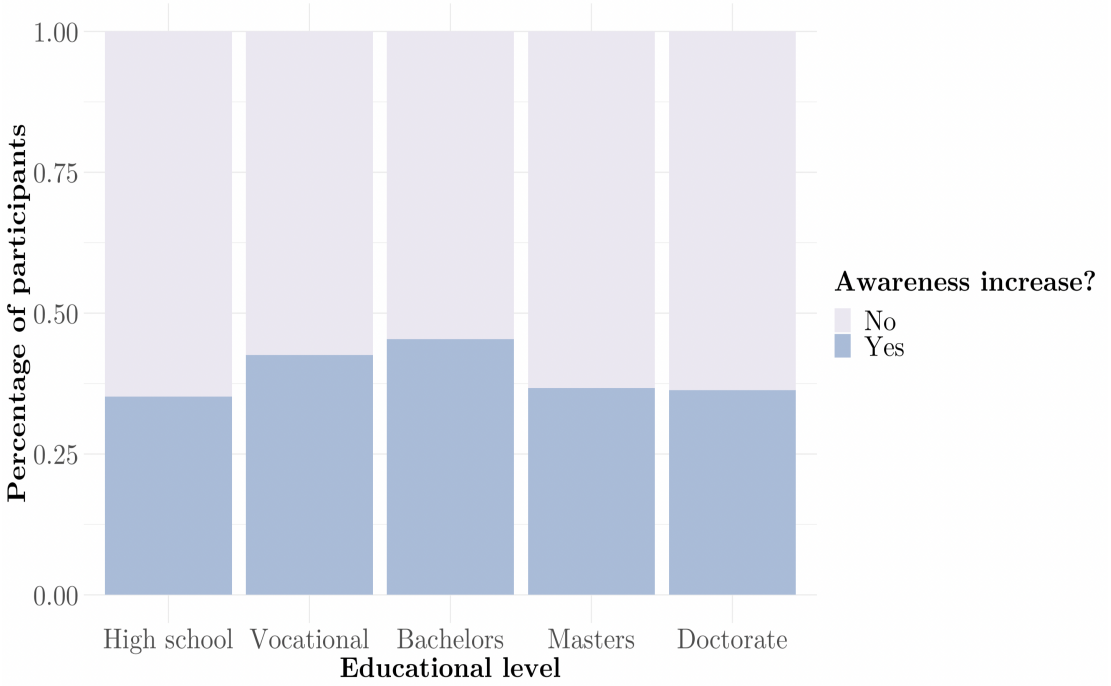

Figure 3: Percentage of participants who exhibited higher awareness after the intervention between education levels

Education had the highest increase in awareness among fields of work, followed closely by education and healthcare. On the other hand, the fields with the lowest increase in awareness were STEM (excluding AI) and law and government. Note that the fields of AI and military were not considered in these conclusions due to the limited number of participants in each (only 5 and 2, respectively). More information can be found in Figure B in the Appendix.

4.2.2 Human Extinction Percentage

The results for the Human Extinction Percentage indicator showed that the mean percentage increased from approximately 13 to 20 percentage between the pre-treatment and

post-treatment results. Additionally, the standard deviation also increased, going from around 18 to around 23. This information can be found in Table 1 below.

Table 1: Descriptive statistics for Human Extinction Percentage indicator

Pre-percentage				Post-percentage			
Min	Max	M	SD	Min	Max	M	SD
0	100	13.53	17.89	0	100	20.34	22.62

There was a rise in the mean percentage of participants' responses for the probability of AI causing human extinction, in both the United States and the Netherlands. The increase in the percentage of respondents was more evident in the United States, where it rose from approximately 14 to 21 percent, compared to the Netherlands, where it increased from around 13 to 18 percent. Figure 4 reveals that the distribution of pre-treatment responses in both countries was comparable, but the post-treatment results in the United States showed a wider range of responses. Prior to the intervention, most answers were within the range of 0 to 25 percent, while post-treatment answers ranged from 0 to 35 percent.

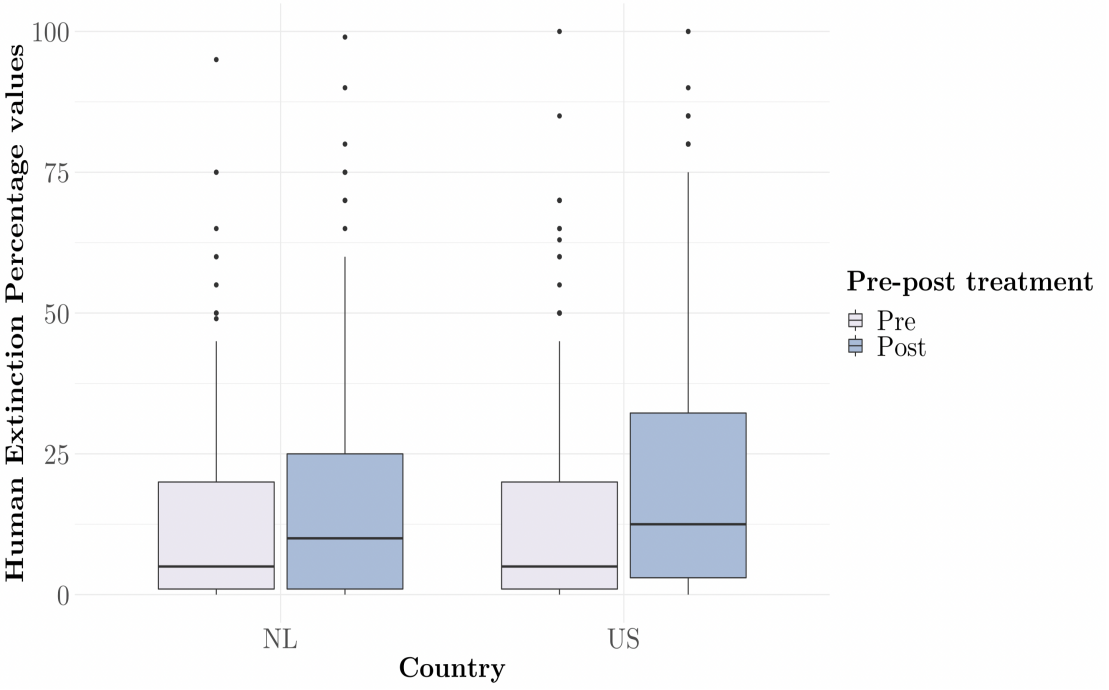

Figure 4: Pre-post summary of the distribution of percentage values for the Human Extinction Percentage indicator across countries

The analysis of the data revealed that among the surveys, the CNN survey had the highest increase in the average percentage of participants' perception of AI causing human extinction, which rose by 12 points. The Elon Musk and PewDiePie surveys also recorded notable increases of around 10 and 9 points respectively. The other surveys showed an increase of between 4 and 7 points, with the Vox survey having the smallest increase at around 4 points. These increases were greater for interventions that also showed greater changes in the level of awareness. These findings suggest that video formats may be more effective than articles in influencing people's views on AI existential risks or that media items featuring well-known figures such as Stephen Hawking, Elon Musk, and Felix Kjellberg may be

more effective in influencing people's perception of AI existential risks. Figure 5 illustrates the range and distribution of participants' responses before and after the intervention for each survey.

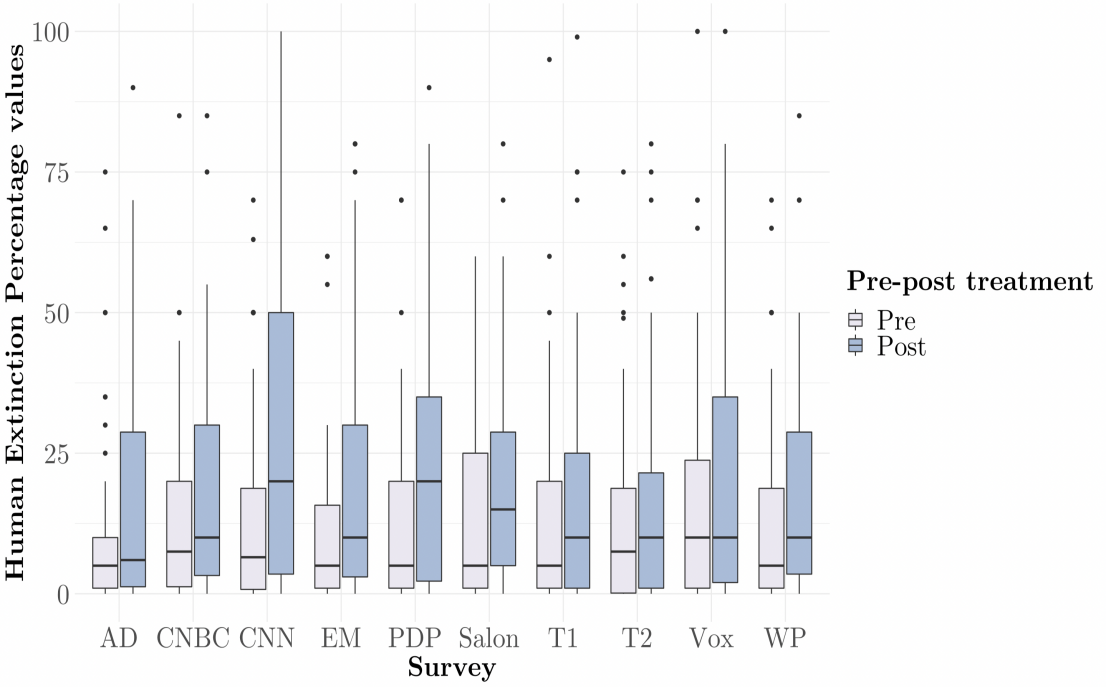

Figure 5: Pre-post summary of the distribution of percentage values for the Human Extinction Percentage indicator across surveys

Table 2 showcases a model of regression analysis that examine the effect of various socio-educational factors on delta percentage. The model considers age, gender, education level, country, media format and field of work. Statistical significance was determined using a p-value of less than 0.05. Model 1 is significant and explains 7 percent of the variance in delta percentage ($R2 = .07$, $F\ (9, 475) = 4.102$, $p < 0.01$).

Table 2: Model 1 - Multivariate analysis of delta percentage with age, education level, country, media format, and work field

	Dependent variable:
	Human Extinction Percentage
	Model 1
Age	0.029
	(0.048)
Female	4.259***
	(1.173)
Bachelors	3.031**
	(1.440)
Doctorate	0.065
	(3.012)
Masters	0.471
	(1.893)
Vocational education	0.011
	(2.218)
United States	0.959
	(1.466)
Video	4.808***
	(1.395)
STEM	−0.277
	(1.536)
Constant	0.452
	(2.157)
Observations	485
R^2	0.072
Adjusted R^2	0.055
Residual Std. Error	12.662 (df = 475)
F Statistic	4.102*** (df = 9; 475)
Note:	*p<0.1; **p<0.05; ***p<0.01

The results of Model 1 indicate that there are significant effects of gender, possession of a bachelor's degree, and video format on the delta percentage, even after accounting for other variables. The gender indicator has male as the baseline category, the education level indicator has high school as the baseline category, the country indicator has the

Netherlands as the baseline category, the media format indicator has article format as the baseline category, and the field of work indicator has "not STEM" as the baseline category.

In Model 1, female gender, possession of a bachelor's degree, and video interventions were found to be positively associated with the delta percentage, with coefficients of (b = +4.26), (b = +3.03), and (b = +4.81), respectively, after controlling for other variables. No significant effect was detected for age or any other education level, country, or field of work. The absence of significance for field of work could indicate that being in STEM may not influence the difference in awareness between genders, given the underrepresentation of women in STEM. In addition, the lack of significance for country may suggest that media format, particularly video format, may be the reason for higher awareness among Unites States participants since all video interventions were targeted towards American participants.

4.3 Views on AI Development and Its Risks

4.3.1 Why Would AI Not Become an Existential Threat?

The survey participants were asked to explain why they believe AI would not be a threat to human extinction both before and after the intervention. Seven predetermined categories, shown in Figure 6, were provided along with a category for custom answers referred to in the figure as "other reasons". Participants could select multiple reasons. In both countries, the most popular reason before the intervention was that AI sounds too science-fictional and unrealistic, while the least popular was the belief that AI could never be that intelligent.

After the intervention, the most common reason in both the Netherlands and the United States became the belief that regulators would prevent AI from reaching a dangerous level, while the least selected reason remained the belief that AI could not become that smart. This may suggest that as participants became more aware of AI's potential dangers, they put more trust in institutions to prevent the technology from becoming an existential threat. This shift in reliance may be more pronounced in the Netherlands.

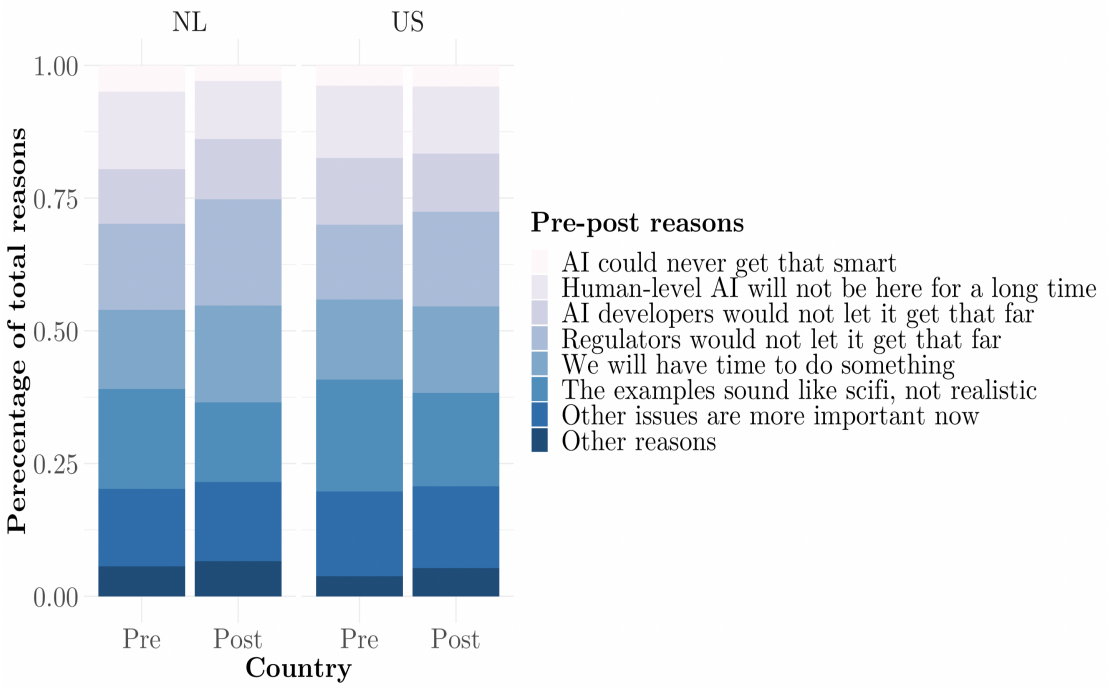

Figure 6: Pre-post percentage of participants' reasons for AI not posing an existential threat across countries

4.3.2 What Is the Role of the Government in AI Development?

The participants were asked for their opinions regarding the appropriate role of the government in AI development. Five predefined categories, as illustrated in Figure 7, were offered with no provision for custom answers. Participants were permitted to select only one option. In both countries, both prior to and following the intervention, the most frequently chosen option among participants was that the government's role should be limited to regulation of AI. After the intervention, in both countries, there was an increase of approximately 3 percent in the proportion of participants who selected the option that the government should prohibit AI. In the United States, there was a decrease in the proportion of participants who believed that the government should refrain from intervening in AI development, from 14 to 16 percent, followed by a smaller decrease in the proportion of participants who believed that the government should regulate and finance AI, of around 1 percent. The United States had a significantly higher proportion of participants who believed that the government should not intervene in pre- and post-intervention, despite a decrease of approximately 4 percent post-intervention. This may indicate that as participants gained a better understanding of the potential hazards posed by AI, some of them became more inclined to support the idea of prohibiting government involvement in AI development.

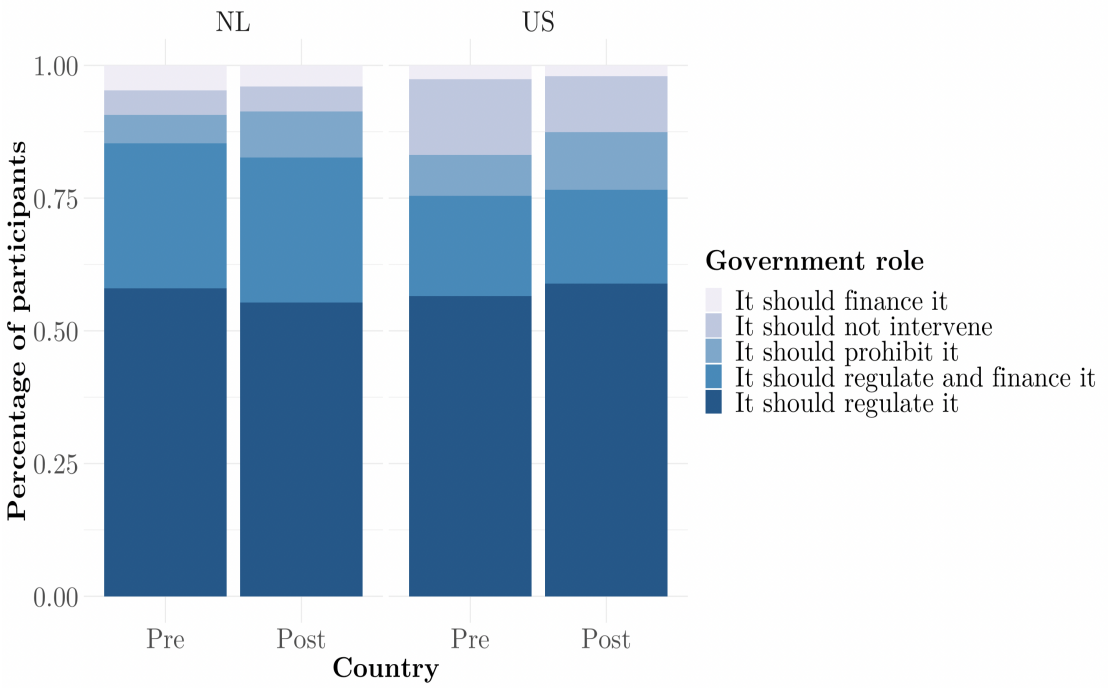

Figure 7: Pre-post percentage of participants' views on the government role in AI development across countries

4.3.3 What Is the Role of the Military and Private Companies?

The participants were asked to share their views on the appropriate role of both military and private companies in AI development. Four predefined categories, as shown in Figure 8, were provided and there was no opportunity for custom answers. Participants were only allowed to select one option. In both countries, the most favoured option among participants prior to the intervention was for both military and private companies to work on AI development. After the intervention, however, the most preferred option shifted towards neither military nor private companies being involved in AI development in the Netherlands, while the United States still maintained its initial position. However, in both

countries there was a decline in the proportion of participants who believed that both military and private companies should work on AI, with a decrease of 5 percent in the Netherlands and 2 percent in the United States. The options for either military or private companies to work on AI were the least favoured, with private companies being two or three times more popular than the military. This may indicate that as participants became more aware of the potential dangers of AI, some became more inclined to support the idea that neither the military nor private companies should work on AI development, with a lesser impact in the United States.

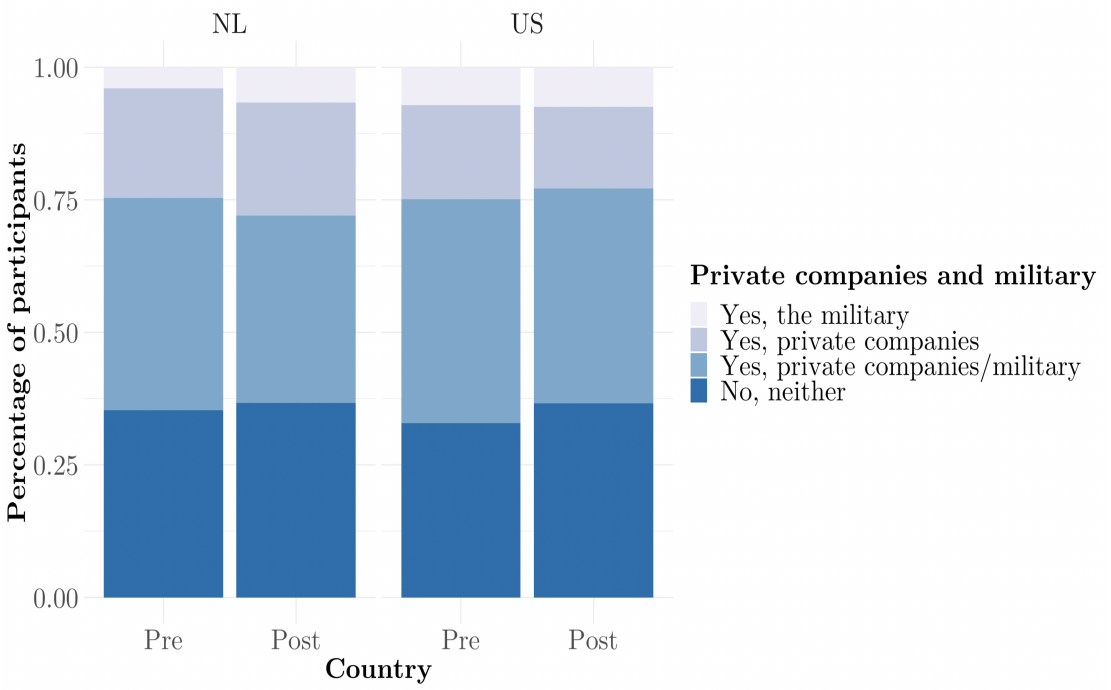

Figure 8: Pre-post percentage of participants' views on the role of private companies and the military in AI development across countries

4.3.4 How Concerned Are Participants About AI Existential Risk?

The participants were questioned on their level of worry regarding the likelihood of human extinction brought about by artificial intelligence. Four predetermined choices, as shown in Figure 9, were provided without the possibility of a custom response, and participants were only allowed to pick one. In both countries, the majority of participants before and after the intervention expressed that they were not concerned about AI causing human extinction. Prior to the intervention, about 68 percent of participants in both countries selected this option, while after the intervention, it was around 52 percent, resulting in a decrease of 16 percent in the number of participants who were not concerned from before to after the intervention. These findings suggest that the media materials had a relevant impact on increasing participants' concern about AI existential risk, which is consistent with the other two key indicators, Human Extinction Events and Human Extinction Percentage.

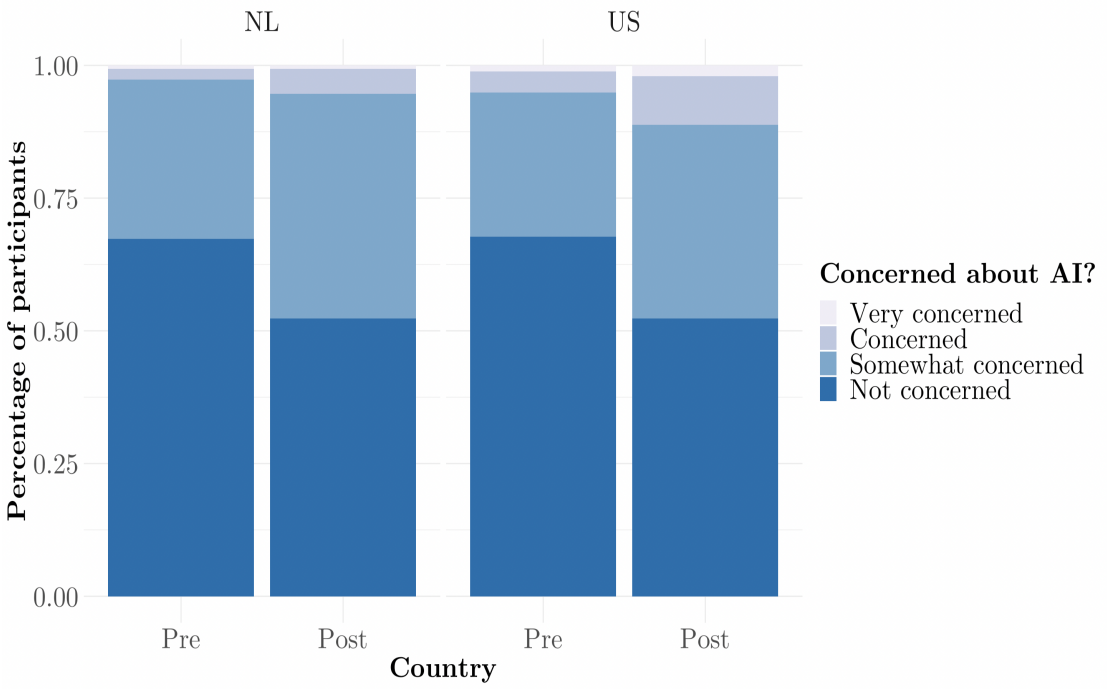

Figure 9: Pre-post percentage of participants' views on their level of concern regarding the existential threat of AI across countries

4.4 Media Trust and Search Engine Preferences

4.4.1 Do Participants Know or Trust the News Channel?

The participants were asked if they knew the news channel they just read the article from. Three predetermined choices, as shown in Figure 10, were provided without the possibility of a custom response, and participants were only allowed to pick one. Mainstream news channels in the US, such as CNN and the Washington Post, had high recognition rates (above 75 percent), while CNBC had a lower rate (64 percent). YouTube videos had lower recognition rates, with PewDiePie performing better at 60 percent compared to Elon Musk's

video at 24 percent. Salon, a more niche US news channel, had a 34 percent recognition rate. Results in the Netherlands were mixed: the well-known newspaper company AD had 78 percent recognition, while Trouw had conflicting results (80 percent and 48 percent). The results indicate that more prominent and established news channels are more widely recognized by participants, which aligns with conventional expectations. However, the recognition PewDiePie, a YouTuber, at 60 percent may come as an unexpected result. Moreover, CNN performed particularly well, with 88 percent recognition from participants, making it the survey with the highest results in both indicators of "Human Extinction Events" and "Human Extinction Percentage." This may indicate that the more well-known a news channel is, the greater its impact on readers. See Figure F in the appendix for more details.

After being asked if they were familiar with the news channel, participants were then questioned about their level of trust in the channel. Mainstream news channels in both countries received higher trust scores, with the one of surveys with an article from Trouw having the highest positive response. Niche or lesser-known news outlets such as YouTube channels and newspapers such Salon and Vox had lower levels of trust. The Elon Musk video received the lowest trust scores, with only 2 percent of participants trusting the source and 62 percent somewhat trusting it. Dutch mainstream news outlets, such as AD and Trouw, received higher trust scores compared to American mainstream channels such as CNN and The Washington Post. This may indicate that Dutch residents have a higher trust in their local newspapers than Americans have in theirs. In addition, the data may

suggest that more established and widely recognized news channels are perceived as more trustworthy than those associated with public figures such as YouTubers.

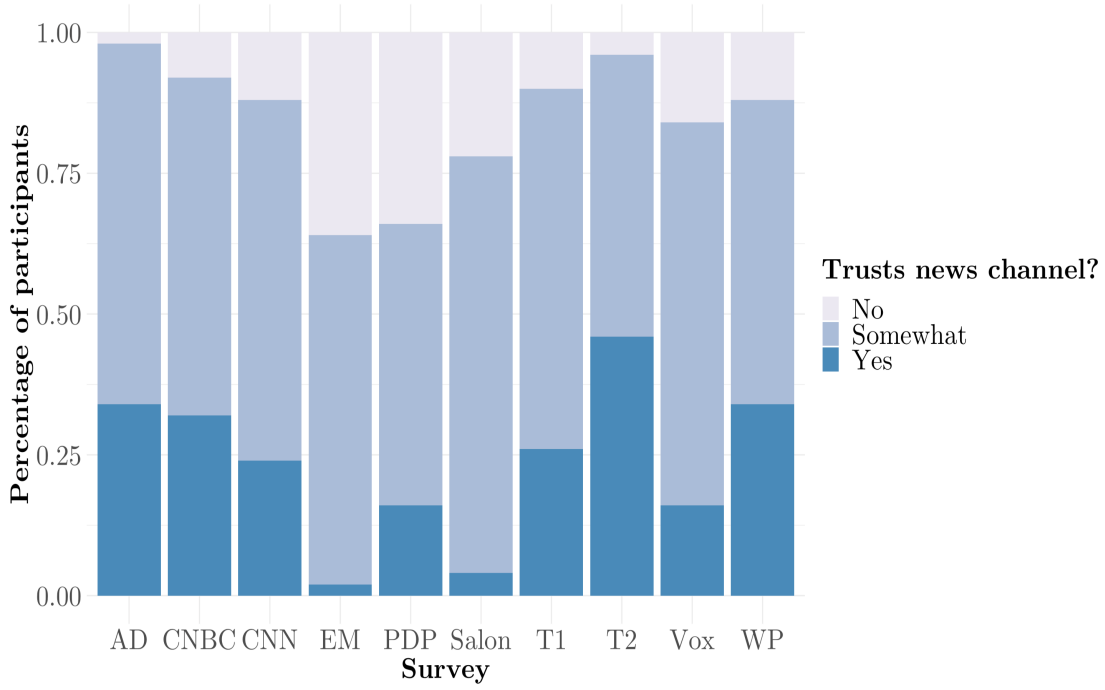

Figure 10: Pre-post percentage of participants' views on their level of trust in the intervention's news channel across countries

4.4.2 Where Would Participants Look For Further Information?

The survey participants were asked about where they would look for information if they wanted to learn more about the topic. Seven predetermined options and a category for custom answers (referred to as "other means") were given, and participants could select multiple sources. In both the United States and the Netherlands, the most popular choice was academic articles at around 35 percent. The second most popular choice was "videos

on YouTube" in the United States and "publications in newspapers" in the Netherlands, with 18 percent and 20 percent respectively. The least preferred choices were "tv series or movies" in both nations, followed by "social media" in the Netherlands at 5 percent and "blog posts" in the US at 5 percent. Wikipedia was also a popular choice for both countries at 12 percent. This may suggest that "videos on YouTube" and "social media" are more popular in the US compared to the Netherlands, who instead chose "publications in newspapers" and "blog posts" more often. This is consistent with the previous conclusion that Dutch citizens may have more trust in their local newspapers than Americans.

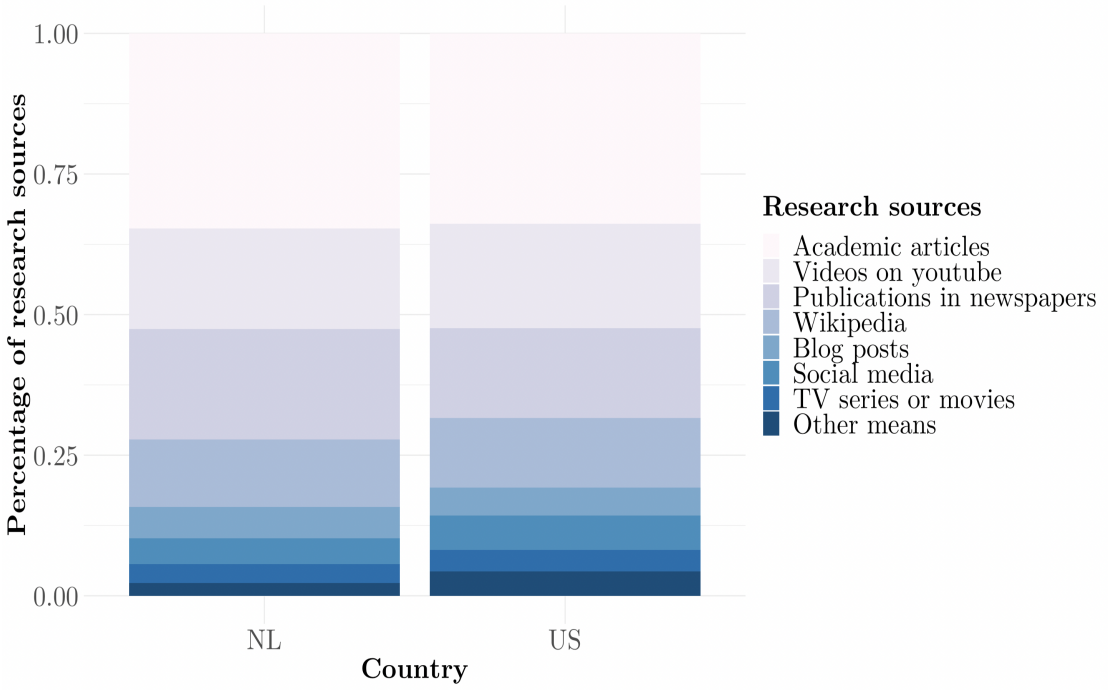

Figure 11: Percentage breakdown of participants' preferred research sources for further researching about AI existential threat across countries

4.5 Motivation for Engagement with AI Post-Research

4.5.1 Does the Information in the Media Item Instigate Further Research?

The survey participants were asked if the information in the article prompted them to do more research on the potential dangers of AI. They were given two options: yes or no. Most participants in both the United States and the Netherlands answered no. The United States had a higher yes response rate of 43 percent, compared to the Netherlands' 35 percent. Among the surveys, the ones with the highest yes response rates were well-known American news companies such as CNN, CNBC, and The Washington Post, with 48, 46, and 44 percent yes responses, respectively. Meanwhile, Dutch news companies like AD and Trouw had lower numbers with yes response rates between 30 and 40 percent. The videos featuring Elon Musk and PewDiePie had overall lower yes response rates compared to the articles, which recorded 40 and 36 positive responses respectively, despite outperforming both Trouw surveys. On the other hand, Stephen Hawking's video on CNN received the most favourable response in the surveys. This may indicate that Americans may be more inclined to undertake further research when provided with information about the existential risk of AI, as compared to individuals residing in the Netherlands.

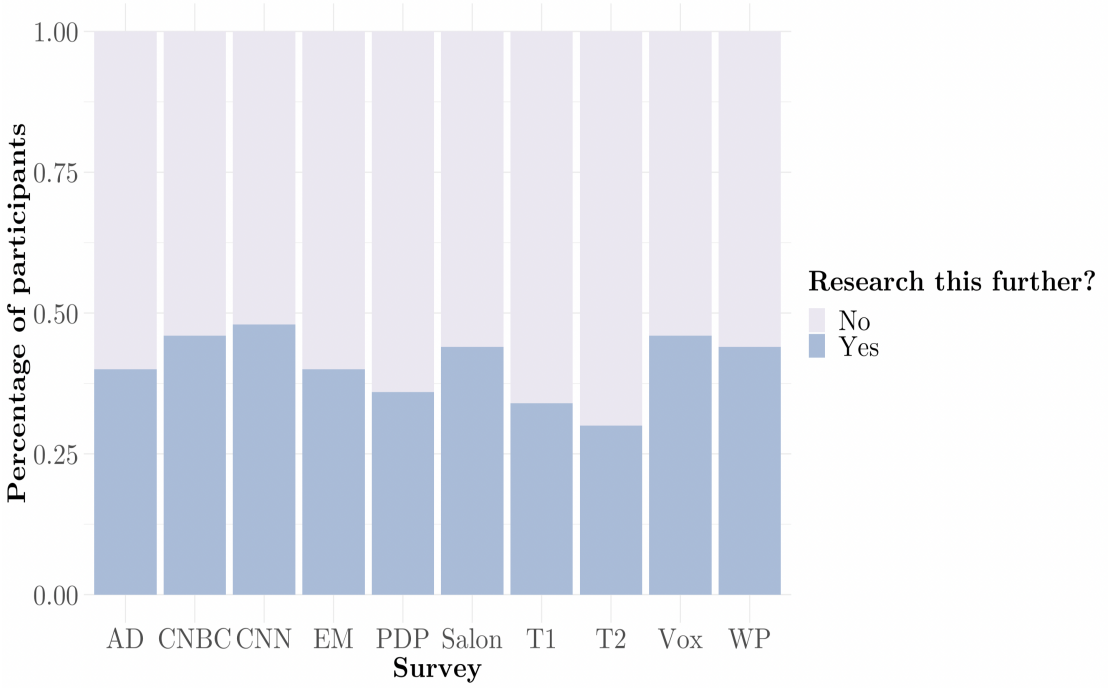

Figure 12: Percentage of participants willing to conduct further research on AI existential threat across surveys

4.5.2 Would Participants Share the Media Content With Friends/Family?

The survey asked if the participants would be willing to share the information in the article with their friends and family, giving them the choice of yes or no. In the Netherlands, 51 percent of participants answered yes while 49 percent answered no. On the other hand, in the United States, the majority of participants answered no with 54 percent opposed to 46 percent who answered yes. The results from different surveys were contrasting, with news channels such as CNN, Vox, AD, and Washington Post having the highest percentage of yes responses at 72, 58, 56, and 56 percent, respectively. Conversely, news channels CNBC, Salon, and videos featuring PewDiePie and Elon Musk had the lowest percentage of yes

responses at 40, 28, 36, and 30 percent, respectively. Trouw had a moderate percentage of yes responses at 52 and 44 percent for each survey. CNN had an exceptionally high percentage of yes responses which correlated with the results of other parameters in the survey.

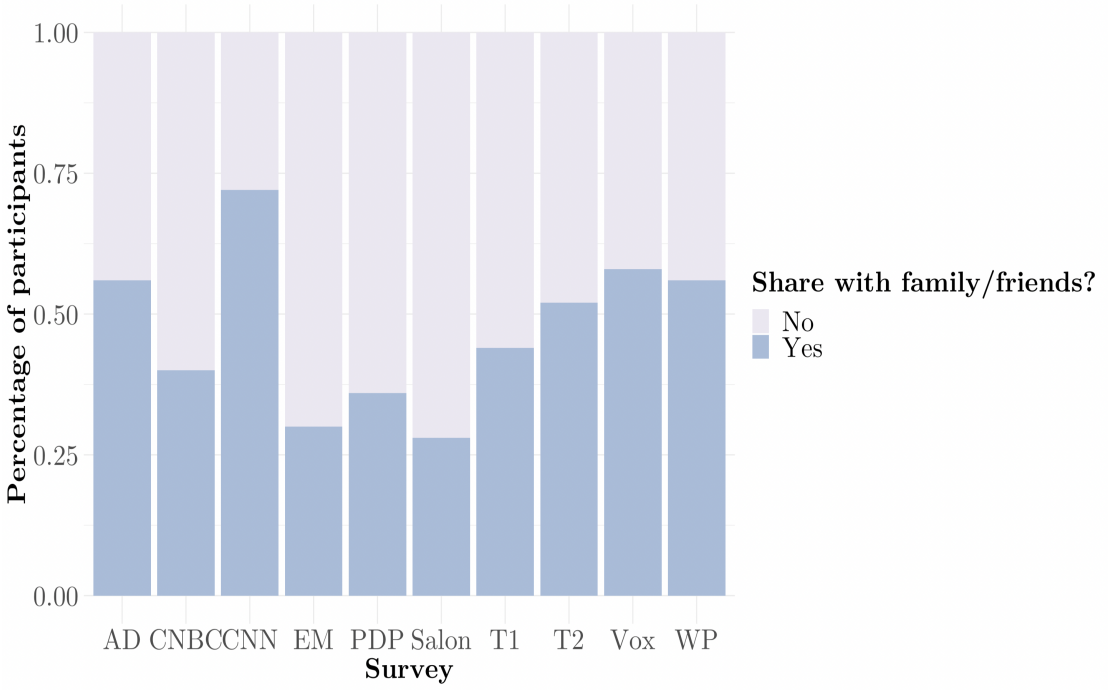

Figure 13: Percentage of participants willing to share the media item information with their friends and family across surveys

4.5.3 Do AI Professionals Want to Participate in AI Safety Development?

The survey asked participants employed in the artificial intelligence sector about their interest in contributing to the enhancement of safety in the development of human-level AI. Participants were presented with four response options: "yes", "no", "maybe", or "I do

not work in the field of AI". Among those who stated to be employed in the field of AI, which constituted 21 percent of participants, 65 participants replied in the affirmative, 14 percent answered negatively, and 21 percent expressed indecision. In both the Netherlands and the United States, the majority of AI-affiliated participants responded positively, with 66 percent and 64 percent respectively. This may indicate that the majority of AI experts in both countries are open to contributing to the improvement of safety measures in the creation of human-level AI, after viewing a media item regarding the AI existential hazards.

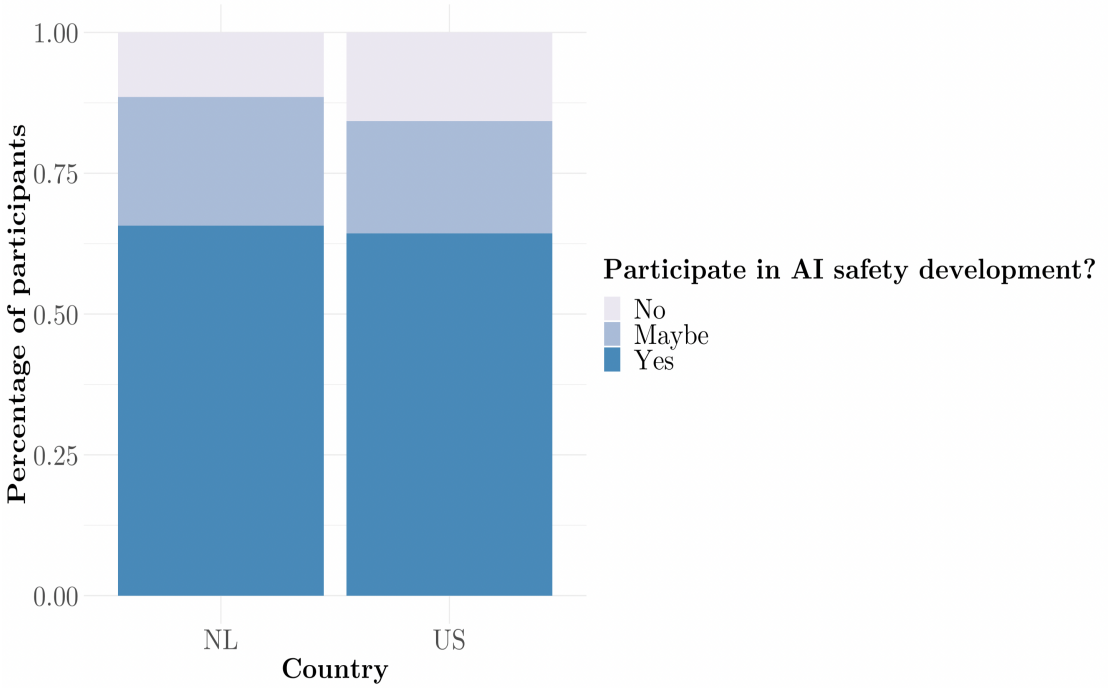

Figure 14: Percentage of AI professionals willing to participate in AI safety development across countries.

4.5.4 Would Participants Volunteer or Donate to AGI Related NGOs?

The survey asked participants their willingness to either volunteer or make a donation to an organisation that aims to educate the public about the potential hazards posed by human-level AI. Five pre-designated categories, as shown in Figure 15, were provided to the participants, along with an option for custom responses, referred to as "other answers." The participants were given the freedom to choose multiple options. For both countries, the results of the survey indicated that a majority of the participants, accounting for 51 percent, expressed a willingness to contribute to these non-governmental organisations in some form. Conversely, 49 percent responded negatively. Nearly half of participants from both the Netherlands and the United States replied negatively, with 51 and 48 percent respectively. The United States had a higher percentage of individuals who expressed a potential interest in volunteering, at 21 percent, compared to 14 percent in the Netherlands. However, the Netherlands recorded a higher percentage of individuals who expressed a potential interest in making a donation, at 24 percent, compared to 19 percent in the United States. These results suggest that patterns of donation and volunteering for non-governmental organisations may differ between the two countries, with possibly a greater focus on volunteering in the United States and a greater focus on donations in the Netherlands.

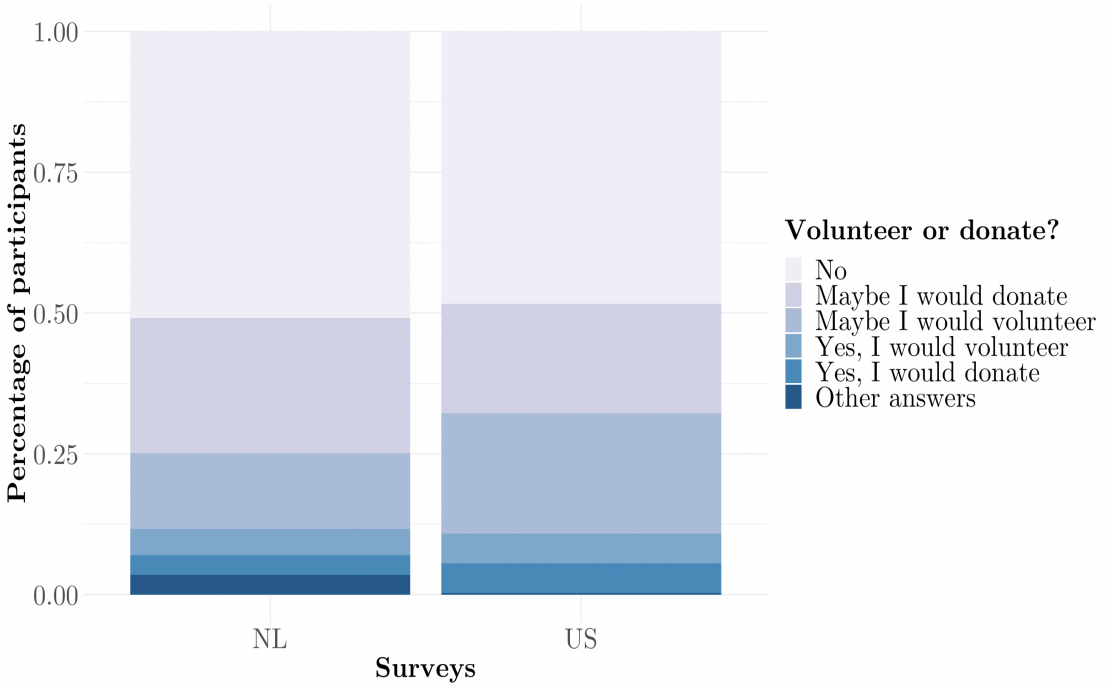

Figure 15: Percentage of participants who are willing to volunteer and/or donate to an organization with the aim of educating people about AI existential threat across countries

4.6 Intersection: Participants' Views and Human Extinction Events

4.6.1 Media Trust and Human Extinction Events

In this segment, the present study analyses the correlation between media credibility and the Human Extinction Event indicator. To this end, a subset of the data was selected that only comprised participants who responded "yes" or "somewhat" to the level of trust they had in the news source from which they had just read an article. This subset was then utilised to evaluate the results from Human Extinction Events indicator from individuals who possess some degree of trust in the media source, which was then compared to the

general dataset, encompassing all participants, including those who did not have faith in the news source.

With respect to conventional mainstream newspapers, there was limited fluctuation in the percentages. The figures for AD increased from 34 to 35 percent, CNBC increased from 40 to 44 percent, CNN increased from 64 to 66 percent, the first survey of Trouw increased from 44 to 49 percent, the second survey of Trouw increased from 34 to 35, Vox increased from 38 to 43 percent and the Washington Post increased from 26 to 30 percent, resulting in an average increase of 3 percent for these established newspapers. On the other hand, niche newspapers such as Salon showed an increase of 9 percent, from 42 to 51 percent. The largest increase was observed among YouTube channels, such as Elon Musk (from 38 to 56 percent) and PewDiePie (from 44 to 52 percent), with an average increase of 13 percent, comprising increases of 18 percent and 8 percent respectively. Among participants who trusted the news source, CNN still held the highest rate of increased awareness. However, after CNN, Elon Musk, PewDiePie, and Salon had the highest rates, which may imply that if their respective audiences trust them, niche news channels and YouTube channels have a higher rate of increased awareness than conventional mainstream channels.

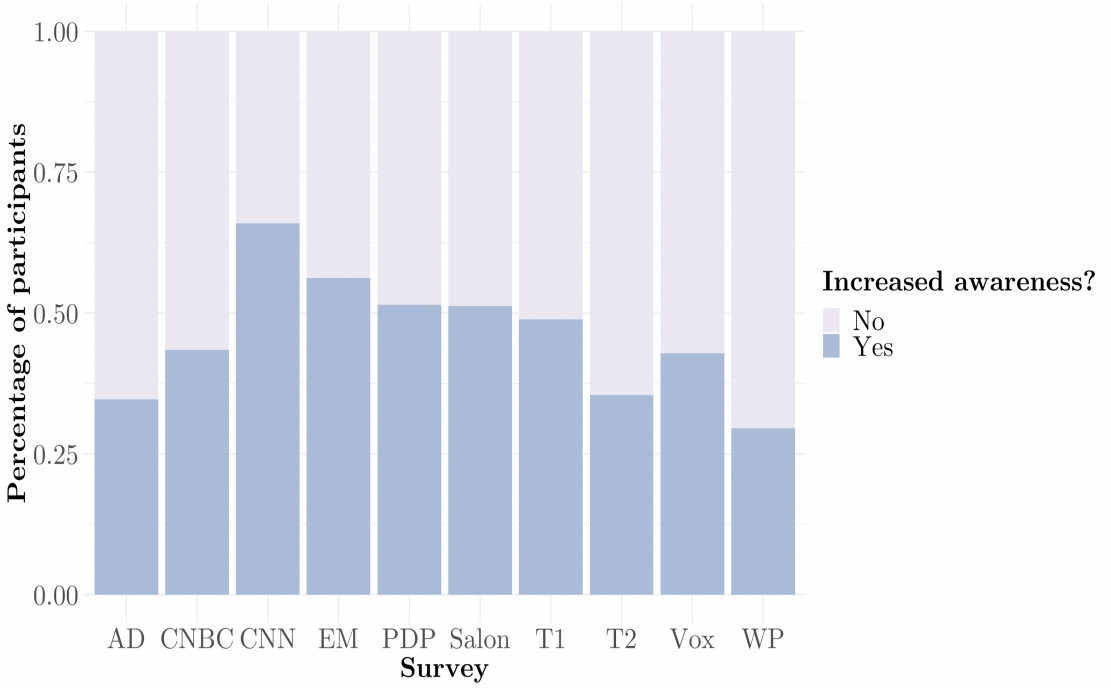

Figure 16: Percentage of participants who reported higher awareness and also affirmed having some level of trust in the news channel across countries

4.6.2 Government Role in AI and Human Extinction Events

In this segment, the research evaluates the correlation between the Human Extinction Events indicator and the participants' perspectives regarding the role of government in AI development. To this end, a subset of the data was selected that only comprised participants who were reported as having their awareness increased in accordance to the Human Extinction Events indicator. This subset was then utilised to evaluate the results from participants perspectives of government role in AI development indicator from individuals who were reported to have been impacted by the media item, showing the difference between pre and post intervention and among the United States and the Netherlands.

The majority in both countries believed that the government should regulate AI, both before and after the intervention. However, there were differences in the changes of public opinion between the two countries. The proportion of participants who believed in government regulation of AI increased in the United States from 58 to 64 percent, while it decreased in the Netherlands from 73 to 70 percent. In addition, before and after the intervention, the proportion of people who believed that the government should regulate and finance AI remained the same in the Netherlands, but decreased from 18 to 12 percent in the United States. The proportion of people who believed that AI should be prohibited increased in both countries, from 5 to 11 percent in the Netherlands and from 11 to 14 percent in the United States. Opinions that AI should be financed were below 3 percent in both countries, both pre- and post-intervention. Furthermore, the proportion of people who believed that the government should not intervene in AI was small in the Netherlands, at around 2 percent, both pre and post-intervention. In contrast, the proportion of such opinion was significantly higher in the United States, at 11 percent before and 8 percent after the intervention.

These results may indicate that there is widespread backing for government regulation of AI in both the United States and the Netherlands among individuals influenced by the media items. Furthermore, the results suggest that there might be a greater emphasis on individual freedom and minimal state intervention in the United States in comparison to the Netherlands. This is due to the fact that a relatively small percentage of people in the Netherlands held the belief that the government should not intervene in AI, while

this proportion was higher in the United States. Lastly, the results show there is limited support for public funding of AI in both countries for those impacted by the intervention, as evidenced by the low percentage of people who believed that AI should be financed.

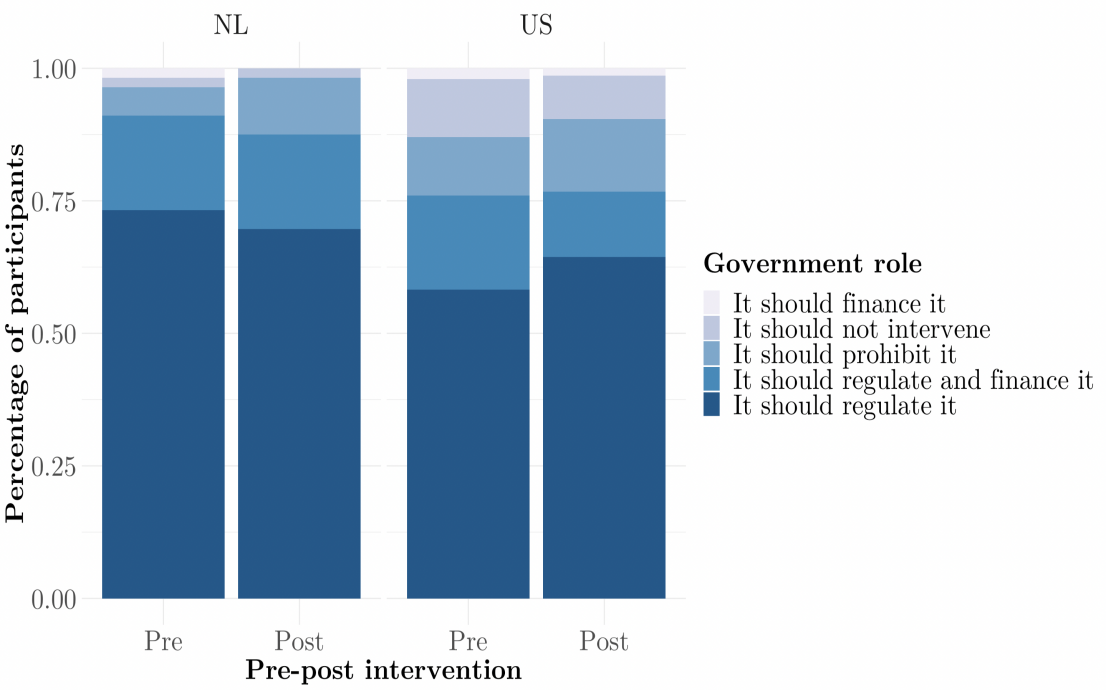

Figure 17: Pre-post percentage of participants who reported higher awareness and their views on the government role in AI development across countries

5 Discussion

5.1 Media Format and Participants Demographics

The findings of this study indicate a consistent pattern where video-based interventions appear to be more effective in increasing awareness than those in the form of articles. This

could be due to the inclusion of prominent figures in the videos, which was demonstrated to have a positive effect. However, it is worth noting that the positive impact of video format cannot be solely attributed to the presence of well-known figures, as both videos and articles included renowned individuals, such as Stephen Hawking, who was featured in both video and article formats on CNN and CNBC surveys respectively.

Furthermore, the results showed that women tend to be more influenced by media interventions compared to men. It was hypothesised that this disparity might be due to the male-dominated nature of the STEM field, where professionals from the field might be more sceptical of claims regarding AI being an existential threat. However, even after controlling for the STEM field indicator, the results indicated that women still had higher levels of increased awareness. Furthermore, the results showed that individuals who possessed a bachelor's degree were associated with greater awareness. It is relevant to note that the sample size of this study was limited (n = 50 per survey), thus the results may not be representative of the general population of these social groups.

5.2 Political Views and Participants Opinions

Three related but distinct patterns emerge from the participants' opinions and political views. Firstly, after being queried as to why they held the belief that AI would not pose an existential threat, the results indicate that upon being presented with information on AI existential risks, the participants later selected the option indicating their belief that regulators would prevent AI from reaching a hazardous level. This may suggest that in the

face of human existential threats, participants tend to view institutions as bastions of safety, manifesting their trust in decision-making bodies to forestall a potential extinction-level event if necessary.

Secondly, as depicted by the results regarding the government's role in AI development, the percentage of participants who believed that the government should either regulate or prohibit AI development increased following the intervention. This may indicate that after being exposed to the media content, a portion of the participants might be prone to believe that the government should regulate or prohibit AI development as a response to the perceived threat, potentially as a result of a fear response to the possibility of extinction.

Thirdly, in a similar vein, the results regarding the involvement of private companies and the military in AI development show that the percentage of participants who believed that neither private companies nor the military should work on AI increased after exposure to media content regarding the potential existential threats posed by AI. This may suggest that after consuming such media content, participants may tend to favour a cessation of AI development by both governmental and private institutions, as previously stated.

It is plausible to assume that educating people about the potential risks of AGI through media interventions may result in them supporting military involvement in AGI development in their country, due to the fear that competitors may develop the technology first. This concern is known as the "information hazard" argument, which has been used to justify not sharing such information to the general public (Bostrom, 2011). However, the empirical evidence of this paper suggests that participants are not more likely to endorse

military involvement in development after being exposed media interventions. The validity of the information hazard argument is called into question since the proportion of people who responded positively to military involvement, and the involvement of both private companies and the military in AI development, did not increase after the intervention.

Nick Bostrom introduces the concept of "attention hazard," which refers to the danger posed by simply drawing attention to specific powerful or relevant ideas or data, even if they are already known (Bostrom, 2011). Bostrom contends that this may make it easier for adversaries to identify potent dangerous areas. For instance, by focusing on the need to prepare for AGI, it might indicate to adversaries that AGI is a promising field to explore (Bostrom, 2011). It is hasty to assume that discussing military involvement in AI becomes innocuous if the public does not support it. Such conversations can generate or exacerbate an attention hazard by increasing the concept's prominence (Bostrom, 2011).

In addition, the results indicate that there is a widespread endorsement of government participation in AI regulation in the Netherlands and the United States among individuals with heightened awareness as per the Human Extinction Events indicator. Among the impacted individuals, most participants in both countries endorsed the belief that the government should regulate AI both before and after the intervention. However, there exists cultural and societal variations in attitudes regarding AI regulation between the two countries. The proportion of individuals who believed that the government should not interfere in AI was greater in the US than in the Netherlands, indicating a higher emphasis on individual autonomy and restricted government intervention in the US.

The results also revealed that the proportion of people who believed that AI should be prohibited increased in both countries, indicating growing concern over the potential risks of AI. Additionally, there was limited support for public investment on AI in both the United States and the Netherlands, with very few people in either country believing that AI should be financed. These results underscore the intricate nature of public attitudes towards AI regulation and the influence of cultural and social factors in shaping these attitudes.

5.3 Media Credibility and Preferences

With respect to media trust and search resource preferences, one anticipated outcome is that individuals are more influenced by well-known news channels, regardless of their own evaluation of the source's trustworthiness, because of a generally perceived higher level of credibility. This can enhance the effectiveness of information dissemination and raise readers' awareness. This study shows that mainstream media channels were more trustworthy to the general population in comparison to YouTube channels. Nevertheless, it is noteworthy that in general Dutch participants exhibited more trust in their local newspapers than Americans.

One notable difference between the United States and the Netherlands is that while scholarly articles were the preferred research resource for respondents in both countries, Americans tended to prefer YouTube videos and social media as secondary sources, while the Dutch tended to to rely more on formal sources. newspaper publications and blog

posts. This might suggest that Americans rely more on informal Internet-based information. However, it is important to recognize the limitations of the study, including a sample size of n = 500 and the possibility that certain behavioural patterns can only be discerned on a larger scale.

While the assertion that mainstream news channels are more credible and thus more effective remains accurate to the general population, a subtlety is evident: when trusted, niche news channels have a higher potential to increase awareness than traditional established channels. In other words, when individuals who read these non-mainstream news channels trust them, they experience a more substantial increase in awareness compared to mainstream channels. This could be because niche news channels attract a specific audience that shares the source's values and ways of disseminating information. For example, to the average person, the information provided by the PewDiePie channel may not be particularly trustworthy, but for fans of the channel who identify with its creator, the conveyed information may hold greater significance than a publication by an anonymous journalist.

5.4 Motivation to Learn about AI Existential Threats

Regarding participant motivation, the data collected in this study suggests that Americans are more inclined than Dutch residents to be motivated to conduct further research on the potential AI existential threats. This may be linked to an existing cultural trend in the United States where conspiracy theories about socio-political issues are widespread. It is

important to note that the potential existential threat of AI is not a conspiracy theory, but due to its seemingly radical nature and lack of discussion in higher decision-making circles, it may be perceived as such. This finding may suggest that Americans are more willing to learn about an issue that is not often addressed in higher political spheres.

Conversely, this study shows that Dutch residents maybe be more likely to consider working in AI safety development. This may be because narrow AI, as opposed to general artificial intelligence, is already widely recognized as being unsafe, such as through biassed algorithms that exclude and harm minorities. One possible explanation for the Dutch tendency to work in safe AI development is the European Union's focus on privacy and security in relation to new technologies. The Union is currently developing the first legal framework for artificial intelligence, which takes a risk-based approach to assessing the safety and legality of AI projects implemented in the European Union territory. However, the small sample size of the Dutch participants and the STEM sample size make it unclear whether this finding is significant or indicative of a general trend.

6 Ethical Considerations

The surveys conducted in the study maintain anonymity by not collecting any personal information such as names, emails, or IP addresses. The questions posed in the survey are general in nature and do not target any specific individual to ensure adherence to the General Data Protection Regulation (GDPR) guidelines, which stipulate that privacy is only violated if the collected data can identify an individual. The survey platform,

Prolific, further enhanced anonymity by assigning each participant a unique ID, which the researchers can have but without access to any personal details provided by the participant to the platform. At start of the survey, the purpose of the study is clearly communicated to the participants, and it is explicitly stated that all collected data will remain anonymous. Participants were also provided with a clear consent form and the option to opt-out at any time.

7 Limitations

This research is subject to certain limitations that need to be acknowledged. Firstly, the study draws conclusions from a sample size of n=50 per article which may not be indicative of the general population's attitudes towards the subject matter. This is particularly true when examining subgroups of participants such as gender, political affiliation, education level or field of work since there is insufficient data available. As a result, it should be noted that the conclusions drawn from this study are limited in their generalizability beyond the sample under investigation. In instances where the phrase "may suggest" is employed, it indicates that the available evidence is not entirely conclusive.

Secondly, there are several biases that may have influenced the results. Participants may have provided answers they believed would please the researcher, particularly if they were being financially compensated for their answers. Furthermore, the sample was limited to individuals who actively sought out surveys on the Prolific platform, which may not accurately represent the general population. Finally, the study's methodology may have

introduced internal biases, such as the phrasing of questions or the structure of the survey.

Finally, the study has limitations in that it did not account for variables that may impact the effectiveness of communication regarding AI existential threats, including social factors such as access to technology and financial circumstances. In addition, the study did not assess the influence of distinct media narratives, such as alarmist or academic language, on the communication of AI existential threats. Consequently, it is relevant to recognize these constraints and exercise caution when interpreting the results and conclusions of this study. Therefore, future research must account for these limitations to obtain a more comprehensive understanding of the effectiveness of communication strategies for AI existential threats.

8 Conclusion

The aim of this study was to examine the influence of communication strategies on the awareness of AI existential risks among the general population in the United States and the Netherlands. It was found that media items are convincing for a large percentage of readers or viewers, namely between 26 and 64 percent, depending on the media item, as measured by the Human Extinction Events indicator. Because of these relatively high numbers, we conclude that awareness can be raised successfully with mass media items for members of the general public. We also find that raising awareness is significantly more successful for female participants ($p < 0.01$), participants with a bachelor's degree ($p < 0.05$), and for video items ($p < 0.01$). No significant effects were detected for age, any education level

other than a bachelor's degree, country, or field of work. Furthermore, the study reveals that even as individuals become increasingly aware of the possible hazards associated with AGI, there is no rise in the proportion of people who endorse military involvement in its development. This finding may challenge the hypothesis that communicating the risks of AGI could constitute an information hazard by increasing the military's proclivity to work on this technology.

Future research should investigate the impact of narratives on media interventions aimed at raising awareness of AI existential threats. Given the high relevance of social media, it is possible that narratives may be a key factor in attracting readers' attention to concise, entertaining, and brief content. Therefore, further investigation is necessary to determine which narratives are most appealing to specific groups and to the general public. In addition, other aspects, such as content length and different formats beyond those explored in this study, such as academic articles, books, and blog posts, should also be researched. Finally, research on other countries could be valuable for both comparative purposes and to identify general trends, such as whether AI existential risk communication brings leads to higher awareness in developed countries compared to developing countries.

9 References

Baum, S. D. (2017). A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3070741

Bostrom, N. (2011). Information hazards: A typology of potential harms from knowledge.

Review of Contemporary Philosophy, 10, 44–79.

Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies (First edition). Oxford University Press.

Crawford, K. (2021). The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press.

Few, R., Brown, K., & Tompkins, E. L. (2007). Public participation and climate change adaptation: Avoiding the illusion of inclusion. Climate Policy, 7(1), 46–59. https://doi.org/10.1080/1469

Fritsche, I., Jonas, E., Kayser, D. N., & Koranyi, N. (2010). Existential threat and compliance with pro-environmental norms. Journal of Environmental Psychology, 30(1), 67–79. https://doi.org/10.1016/j.jenvp.2009.08.007

Galanos, V. (2019). Exploring expanding expertise: Artificial intelligence as an existential threat and the role of prestigious commentators, 2014–2018. Technology Analysis & Strategic Management, 31(4), 421–432. https://doi.org/10.1080/09537325.2018.1518521

Helmke, G., & Levitsky, S. (2004). Informal Institutions and Comparative Politics: A Research Agenda. Perspectives on Politics, 2(4), 725–740. https://doi.org/10.1017/S1537592704040472

Khatibi, F. S., Dedekorkut-Howes, A., Howes, M., & Torabi, E. (2021). Can public awareness, knowledge and engagement improve climate change adaptation policies? Discover Sustainability, 2(1), 18. https://doi.org/10.1007/s43621-021-00024-z

Lee, T. M., Markowitz, E. M., Howe, P. D., Ko, C.-Y., & Leiserowitz, A. A. (2015). Predictors of public climate change awareness and risk perception around the world. Nature Climate Change, 5(11), Article 11. https://doi.org/10.1038/nclimate2728

Luckerson, V. (2014, December 2). 5 Very Smart People Who Think Artificial Intelligence Could Bring the Apocalypse. Time. https://time.com/3614349/artificial-intelligence-singularity-stephen-hawking-elon-musk/

Ord, T. (2020). The precipice: Existential risk and the future of humanity. Hachette Books.

Pandve, H., Fernandez, K., Khismatrao, D., Chawla, P., Singru, S., & Pawar, S. (2011). Assessment of awareness regarding climate change in an urban community. Indian Journal of Occupational and Environmental Medicine, 15(3), 109. https://doi.org/10.4103/0019-5278.93200

Ramamoorthy, A., & Yampolskiy, R. (2018). Beyond Mad?: The Race for Artificial General Intelligence. 1, 8.

Roose, K. (2022, August 24). We Need to Talk About How Good A.I. Is Getting. The New York Times. https://www.nytimes.com/2022/08/24/technology/ai-technology-progress.html

Russell, S. J. (2019a). Human compatible: Artificial intelligence and the problem of control. Viking.

Russell, S. J. (2019b, October 8). Opinion | How to Stop Superhuman A.I. Before It Stops Us. The New York Times. https://www.nytimes.com/2019/10/08/opinion/artificial-intelligence.html

Shermer, M. (2017). Why Artificial Intelligence is Not an Existential Threat. Skeptic, 22(2), 29–35.

Singh, S. (2021). Crisis Response Framework and Public Policy Response. Indian Journal

of Public Administration, 67(3), 396–412. https://doi.org/10.1177/00195561211032962

Steffen, W., Richardson, K., Rockström, J., Cornell, S. E., Fetzer, I., Bennett, E. M., Biggs, R., Carpenter, S. R., de Vries, W., de Wit, C. A., Folke, C., Gerten, D., Heinke, J., Mace, G. M., Persson, L. M., Ramanathan, V., Reyers, B., & Sörlin, S. (2015). Planetary boundaries: Guiding human development on a changing planet. Science, 347(6223), 1259855. https://doi.org/10.1126/science.1259855

Stollberg, J., & Jonas, E. (2021). Existential threat as a challenge for individual and collective engagement: Climate change and the motivation to act. Current Opinion in Psychology, 42, 145–150. https://doi.org/10.1016/j.copsyc.2021.10.004

Times, T. N. Y. (2014, September 8). Best Selling Science Books. The New York Times. https://www.nytimes.com/2014/09/09/science/best-selling-science-books.html

UN Climate Change, U. (2015, January 15). Scientists Say Planetary Boundaries Crossed. https://unfccc.int/news/scientists-say-planetary-boundaries-crossed

UN News. (2021, October 25). COP26: Praise for updated national climate plans, but 'nowhere near' goal | UN News. https://news.un.org/en/story/2021/10/1103972

Warner, H. (2021). Communication of Existential Risks to Dutch General Public. Existential Risk Observatory.

Wernli, D., Clausin, M., Antulov-Fantulin, N., Berezowski, J., Biller-Andorno, N., Blanchet, K., Böttcher, L., Burton-Jeangros, C., Escher, G., Flahault, A., Fukuda, K., Helbing, D., Jaffé, P. D., Jørgensen, P. S., Kaspiarovich, Y., Krishnakumar, J., Lawrence, R. J., Lee, K., Léger, A., ... Young, O. (2021). Building a multisystemic understanding

of societal resilience to the COVID-19 pandemic. BMJ Global Health, 6(7), e006794. https://doi.org/10.1136/bmjgh-2021-006794

Yampolskiy, R. V. (2016). Artificial superintelligence: A futuristic approach. Taylor & Francis, CRC Press.

A Appendix

A.1 Media items, questionnaire and survey structure

This section provides detailed information about the media items used as intervention in the surveys. It also explains the targeted audience and format type of each media item. Furthermore, it includes a list of online links that were used in the surveys and offered to the participants. The media items used in the research were carefully selected based on their relevance to the research question and their potential to elicit informative responses from the participants. These media items could be in the form of either written publications or videos, and were designed to target specific the general Dutch and American population.

In addition, this section provides an overview of the questionnaire used for data collection, including their ordering and how they fit into the structure of the survey. The survey was structured to have a section on participants' personal information, two sections for measurements (pre- and post-intervention), and a final section with additional questions. Table 3, included in this section, shows how each of the questions used in the survey was intended to fit these categories. This information is relevant for understanding how the data was collected and how the questions were designed to address the research question.

Online links to the media items:

American Public:

- Elon Musk (video): "I Tried To Warn You" - Elon Musk LAST WARNING (2023)

- PewDiePie (video): WE ARE ALL GOING TO D1E. (AI)

- CNN (video): Hawking: A.I. could be end of human race | CNN

- CNBC (publication): Stephen Hawking says A.I. could be 'worst event in the history of our civilization'

- Salon (publication): Human-level AI is a giant risk. Why are we entrusting its development to tech CEOs? | Salon.com

- Vox (publication): The case that AI threatens humanity, explained in 500 words - Vox

- Washington Post (publication): Opinion | What is the worst-case AI scenario? Human extinction. - The Washington Post

Dutch Public:

- AD (publication): Oud-topman Google waarschuwt nu voor de gevaren van kunstmatige intelligentie | Tech | AD.nl

- Trouw 1 (publication): Denk na over kunstmatige intelligentie, want die gaat onze toekomst bepalen

- Trouw 2 (publication): Gelet op de gevaren verdient de digitale agenda in Den Haag meer ambitie

Table 3: Survey sections and questions

Section	Questions
Section 1:	1. How old are you?
Personal information	2. What is your gender?
	3. What is your highest level of education?
	4. What is your field of work?
Section 2 & 4:	5. List three events, in order of probability (from most to least probable),
Pre-post measurements	that you believe could potentially cause human extinction within the next 100 years.
	6. In your opinion, how likely is human extinction to be caused
	by artificial intelligence in the next 100 years in percentage?
	7. Why would you say that AI might not pose a risk of human extinction?
	8. On the development of human-level artificial intelligence,
	would you consider that the government should have a role to prohibit,
	regulate or finance its progress or should it not have that role?
	9. Would you consider that private companies / the military should work
	on the development of human-level artificial intelligence?
	10. Give the main reason why you decided on this answer.
	11. How concerned are you about human extinction caused by artificial intelligence?
Section 4:	12. Do you know the newspaper you just read the article from?
Additional information	13. Do you trust the information provided by this newspaper?
	14. Does the information in this article motivate you
	to do more research on the existential risks of artificial intelligence?
	15. If you wanted to further inform yourself on this topic,
	where would you look for information?
	16. Would you share this information with friends and family?
	17. If you work in the field of Artificial Intelligence,
	would you be interested in working to increase safety
	in the development of human-level artificial intelligence?
	18. Would you volunteer or donate to an organization that educates the public
	about the potential dangers of human-level artificial intelligence?

A.2 Further details about participants: age, education and work

This section delves further into the demographics of the participants involved in the study. Figure A illustrates the distribution of participant ages between surveys, providing a visual representation of the age groups that were included in the research. In addition, Table 4 provides a detailed breakdown of the percentages of participants' educational levels and professional fields among the surveys.

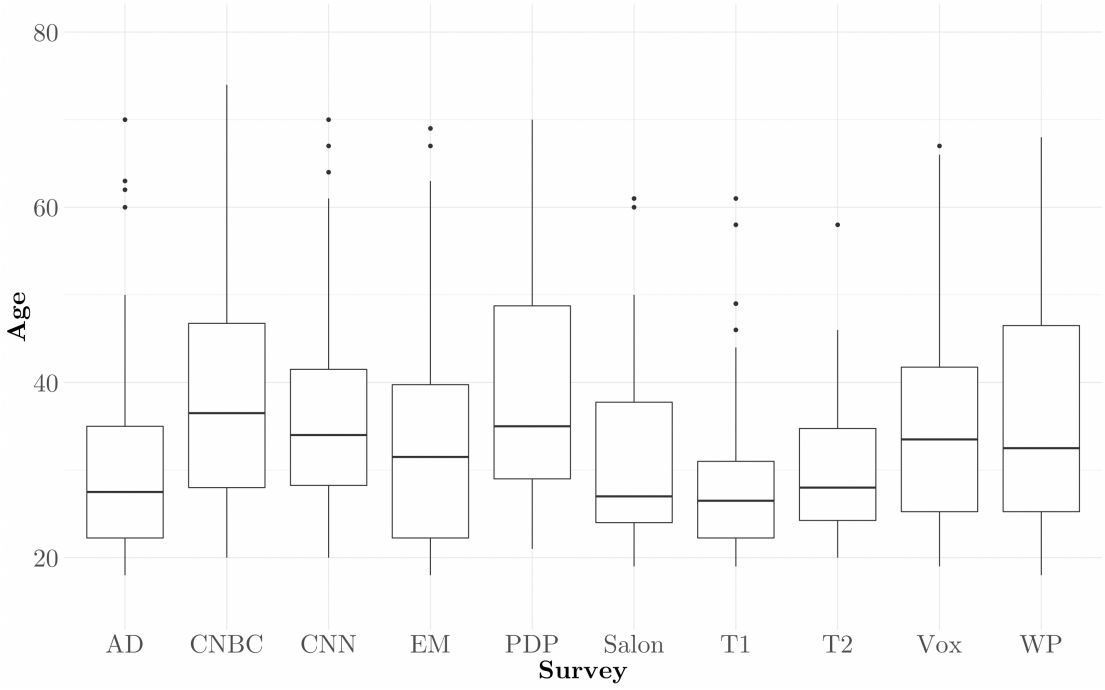

Figure 18: Distribution of participants' ages across the surveys

Table 1: Description of participants' educational background and field of work

	Surveys									
	AD	T1	T2	CNBC	CNN	EM	PDP	Salon	Vox	WP
Education										
High school	8 (16%)	8 (16%)	4 (8%)	16 (32%)	12 (24%)	21 (42%)	20 (40%)	19 (38%)	23 (46%)	14 (28%)
Vocational education	6 (12%)	1 (2%)	2 (4%)	5 (10%)	8 (16%)	8 (16%)	2 (4%)	7 (14%)	3 (6%)	5 (10%)
Bachelors	25 (50%)	28 (56%)	25 (50%)	19 (38%)	22 (44%)	14 (28%)	17 (34%)	15 (30%)	20 (40%)	22 (44%)
Masters	9 (18%)	10 (20%)	18 (36%)	8 (16%)	4 (8%)	3 (6%)	9 (18%)	7 (14%)	2 (4%)	9 (18%)
Doctorate	2 (4%)	3 (6%)	1 (2%)	2 (4%)	4 (8%)	4 (8%)	2 (4%)	2 (4%)	2 (4%)	0 (0%)
Field of work										
Artificial Intelligence (AI)	1 (2%)	0 (0%)	0 (0%)	2 (4%)	1 (2%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (2%)
STEM (not AI)	8 (16%)	21 (42%)	12 (24%)	4 (8%)	9 (18%)	8 (16%)	5 (10%)	11 (22%)	5 (10%)	7 (14%)
Arts and media	6 (12%)	2 (4%)	4 (8%)	6 (12%)	3 (6%)	7 (14%)	3 (6%)	2 (4%)	1 (2%)	5 (10%)
Business and finance	6 (12%)	5 (10%)	3 (6%)	1 (2%)	4 (8%)	6 (12%)	3 (6%)	6 (12%)	3 (6%)	2 (4%)
Education	4 (8%)	2 (4%)	7 (14%)	4 (8%)	3 (6%)	5 (10%)	5 (10%)	4 (8%)	2 (4%)	5 (10%)
Healthcare	1 (2%)	1 (2%)	5 (10%)	4 (8%)	4 (8%)	3 (6%)	4 (8%)	7 (14%)	8 (16%)	8 (16%)
Law and government	4 (8%)	2 (4%)	4 (8%)	0 (0%)	1 (2%)	3 (6%)	0 (0%)	2 (4%)	3 (6%)	1 (2%)
Military	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (2%)	0 (0%)	0 (0%)	1 (2%)
Other	20 (40%)	17 (34%)	15 (30%)	29 (58%)	25 (50%)	18 (36%)	29 (58%)	18 (36%)	28 (56%)	20 (40%)

A.3 Further information about the results obtained

In this section of the research paper, further information regarding the obtained results is presented. This section includes alternative graphs that were not included in the main text but are still valuable in providing a different perspective on the interplay between variables in the study. These alternative graphs provide with additional insights into the results obtained in the study. By presenting different interplays between variables, one can gain a more comprehensive understanding of the relationships between different factors and how they contribute to the overall findings.

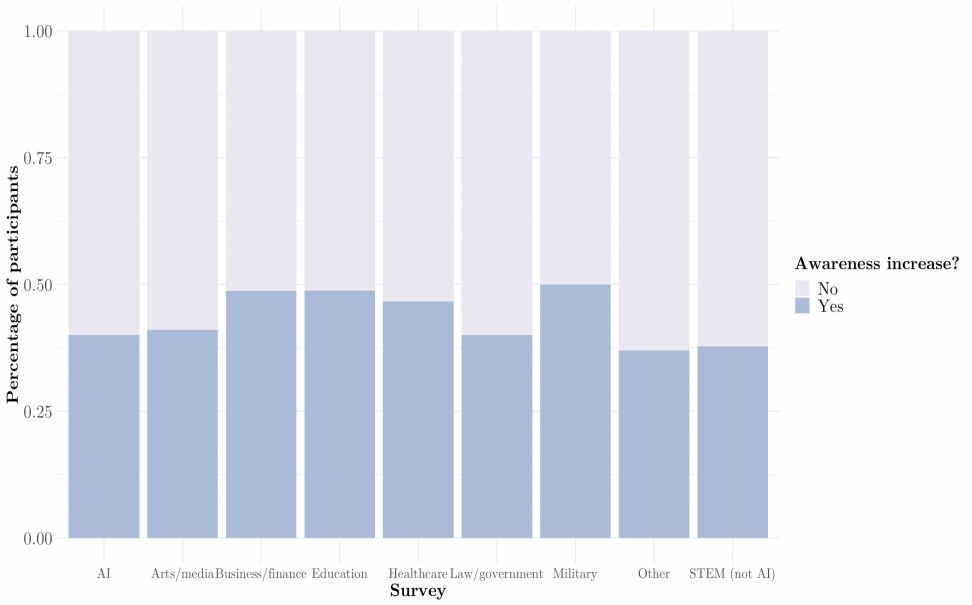

Figure 19: Percentage who exhibited higher awareness after the intervention across professional fields

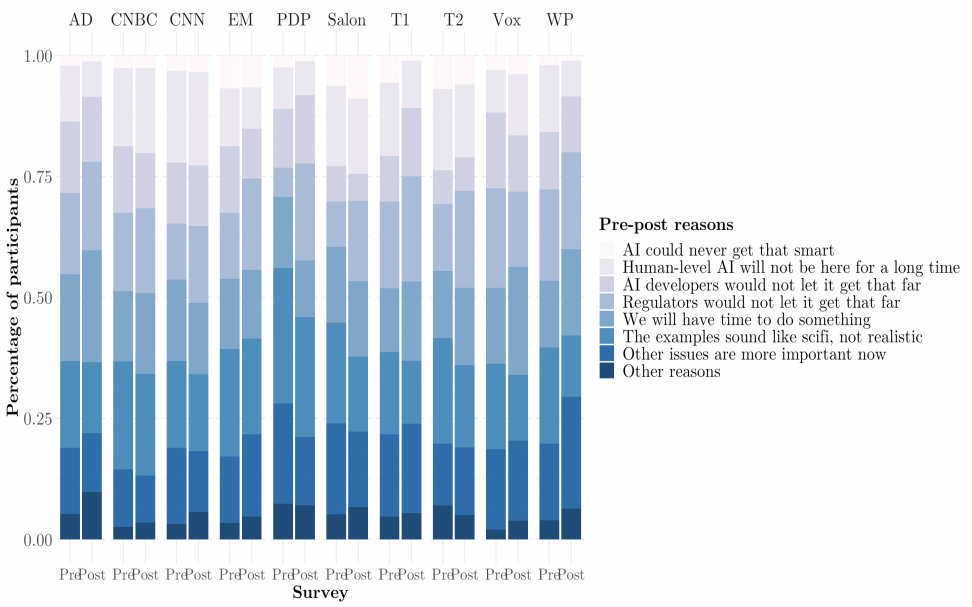

Figure 20: Pre-post percentage of participants' reasons for AI not posing an existential threat across surveys

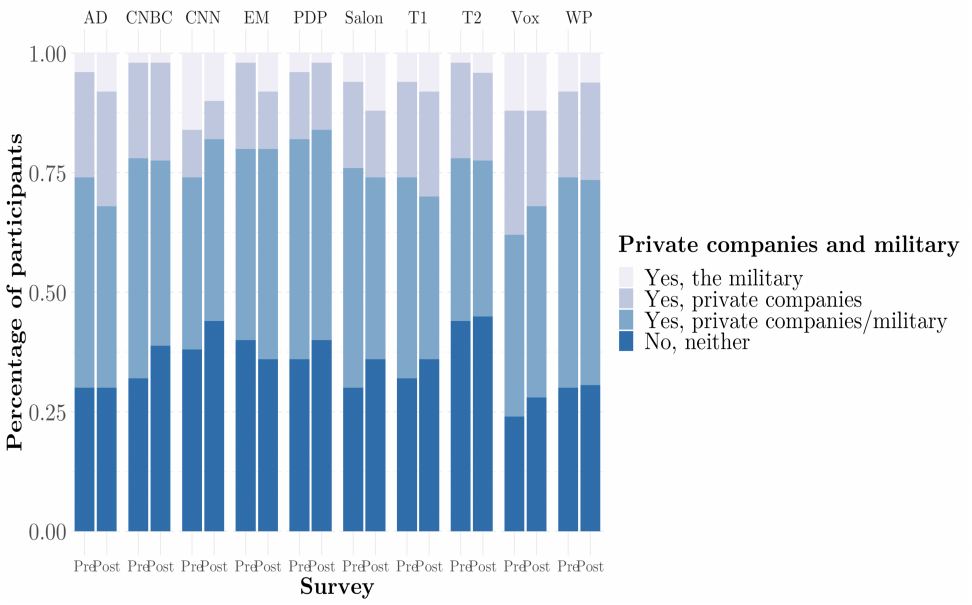

Figure 21: Pre-post percentage of participants' views on the role of private companies and the military in AI development across surveys

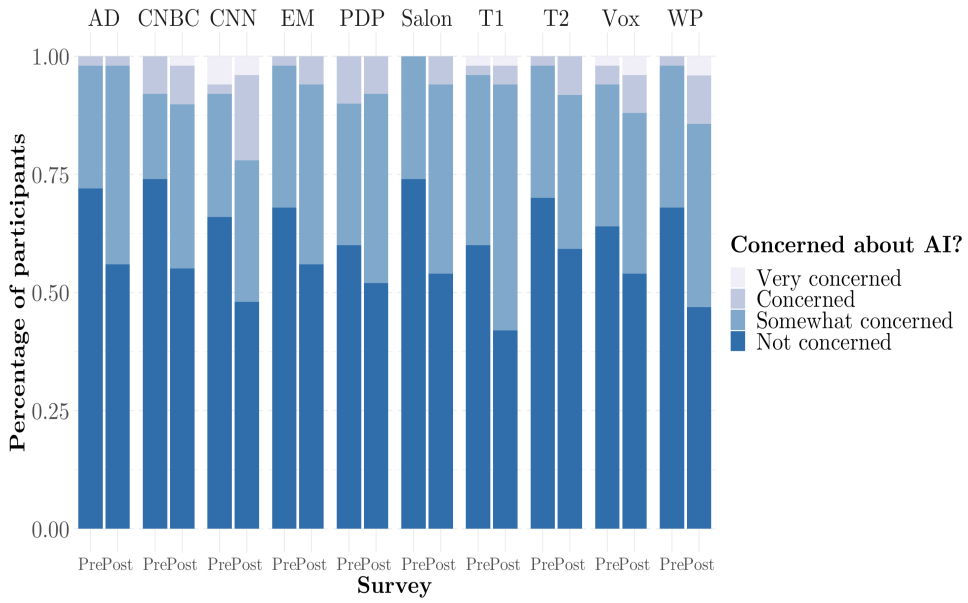

Figure 22: Pre-post percentage of participants' views on their level of concern regarding the existential threat of AI across surveys

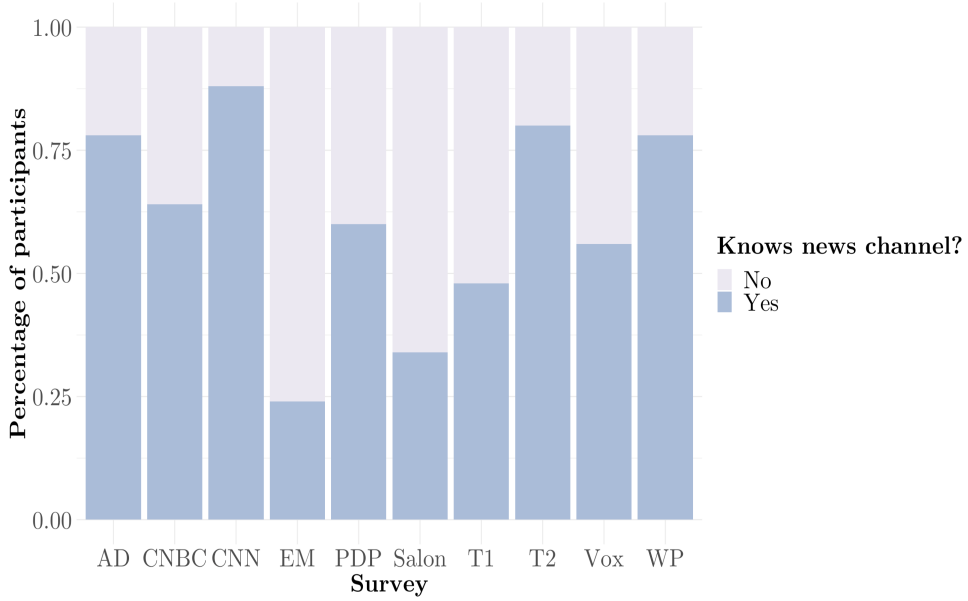

Figure 23: Pre-post percentage of participants' who knew the intervention's news channel across surveys

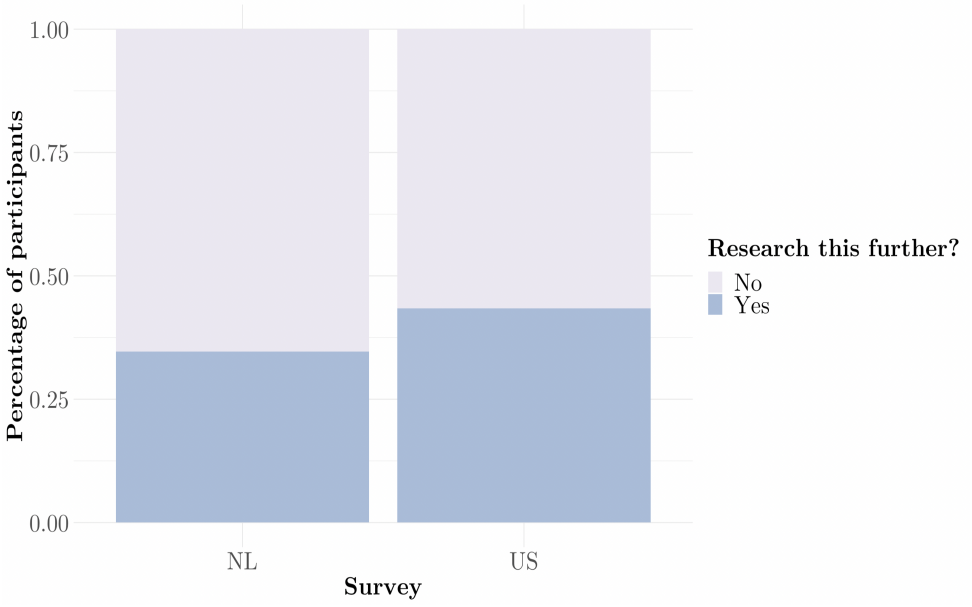

Figure 24: Percentage of participants willing to conduct further research on AI existential threat across countries

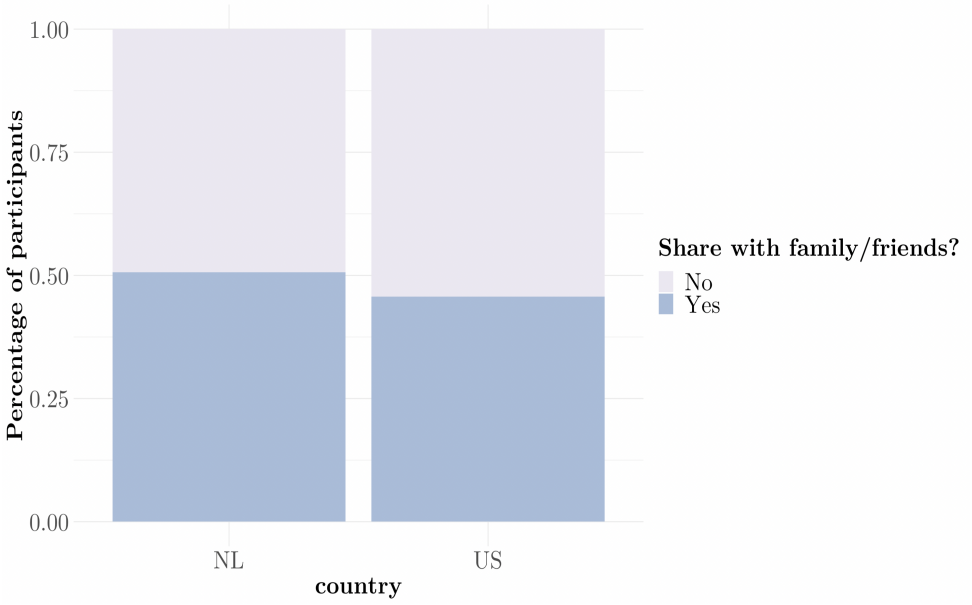

Figure 25: Percentage of participants willing to share the media item information with their friends and family across countries

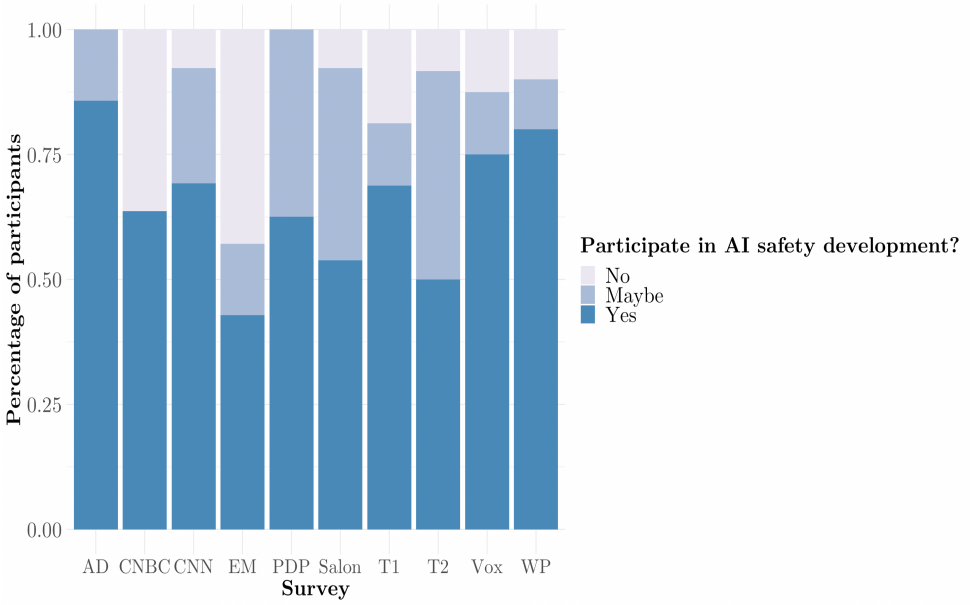

Figure 26: Percentage of AI professionals willing to participate in AI safety development across surveys

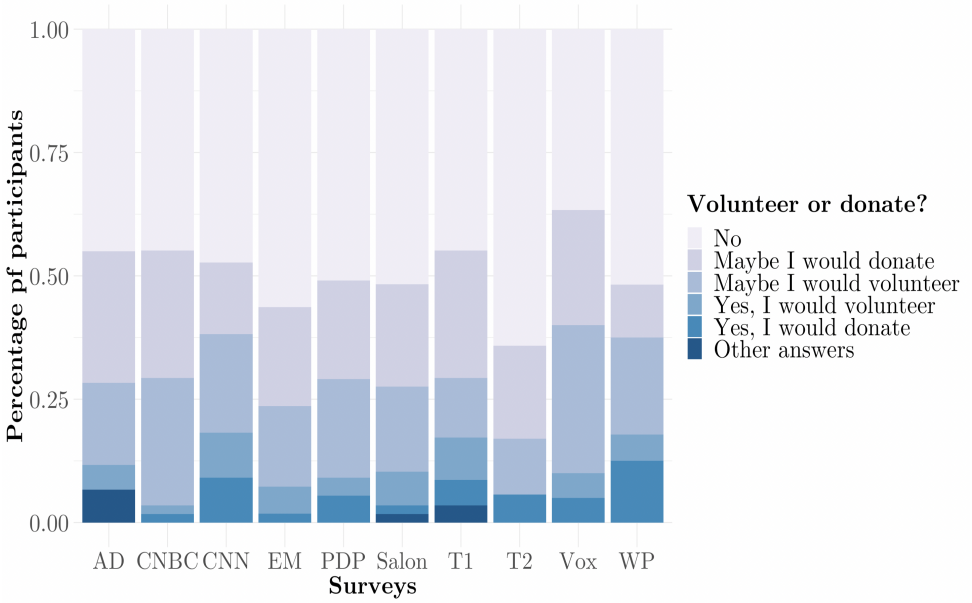

Figure 27: Percentage of participants who are willing to volunteer and/or donate to an organization with the aim of educating people about AI existential threat across surveys