Existential Risks: Who knows? Who cares?

Imogen Gaskell

Laidlaw Undergraduate Research and Leadership Programme

University of York, Department of Psychology

Research Supervisor: Dr Rob Jenkins

Word Count: 2702

Acknowledgements

This project was completed thanks to the Laidlaw Undergraduate Research & Leadership Programme at the University of York. I would like to acknowledge my heartfelt gratitude to Dr. Rob Jenkins for acting as my supervisor and Rana Qarooni and Scarlett Syme for their advice and invaluable guidance throughout the research project. I would also like to extend my thanks to Lord Laidlaw for providing the funding and support throughout the project.

Abstract

When it comes to human extinction, experts suggest the current century is the riskiest yet. However, a relatively small amount of resources have been invested into reducing existential risks and an even smaller amount of research into public awareness. This study aims; (i) to give insight into which risks are seen to be the greatest threat through the ITN framework and worry by the general public; (ii) to compare the results with experts' forecasts on the likelihood of these risks occurring (Ord, 2020); (iii) to have both cognitive and metacognitive measures (respondents' estimates of other people's judgements) for the same survey items. In a quantitative non-experimental correlational design, 150 participants took part in a survey requiring the ranking of existential risks based on the importance, tractability, neglectedness (ITN) framework and worry. Results suggested that climate change, nuclear war and naturally occurring pandemics were considered to be the most worrisome and important. There was a significant difference in ranks between participants and expert predictions, and the metacognitive results suggested egocentric bias and the better-than-average effect had occurred. These findings amplify the uncertainty and disagreement as to where our focus should be held when trying to mitigate these risks within the general public.

Introduction:

For decades, the possibility of human extinction has been the focus of science fiction tropes. However, in the last twenty years, there has been substantial growth in research on existential risk (ERs), as experts suggest that we are at a pivotal moment in deciding on the future of humanity (Bostrom, 2013). Defined by Toby Ord (2015), Existential Risks (ERs) are any risks that have the possibility of ending humanity. Whilst small institutes have begun to emerge, such as the University of Oxford's "Future of Humanity Institute", that are studying these risks and how to mitigate them, it still remains that very little is known about the general public's views on ERs.

The advancement in research about the topic has occurred with the progression in technology. Before the industrial revolution, the biggest threat to humanity's survival was natural risks, such as asteroid impacts, supervolcanic eruptions and stellar explosions (Bostrom, 2013). While these have always posed a risk, humanity has avoided them for hundreds of thousands of years, and the probability of them occurring during the next 100 years is low. However, over the last century, with advancements in technology, anthropogenic (human-driven) risks have begun to raise anxieties. These consist of risks such as climate change, nuclear weapons and unaligned AI. Ord's predictions suggest that unaligned AI has a 10% chance of ending humanity within the next 100 years.

Due to the deadliest risks being anthropogenic, humans have the ability to avoid them. It makes it vital to understand what the general public knows about ERs and what they consider to be most worrisome, as worry has been shown to act as a motivator when enacting change (Goldberg et al., 2021). However, so far (June, 2022) no research has been carried out investigating which risks the general public would prioritise to avoid.

One influential framework that has been used when discussing risk prioritisation is the importance, tractability and neglectedness (ITN) framework. Importance refers to the size of the problem (e.g. the probability it will occur). Tractability refers to how easily in this case the risk can be controlled. Finally, neglectedness refers to how much the risk is neglected versus oversaturated (Gainsburg et al., 2021). In using this framework, it helps to identify what the general public prioritises in terms of ERs.

In this investigation, we aim: (i) to give insight into which risks are seen to be the greatest through the ITN framework and worry by the general public. (ii) to compare the results with experts' forecasts on the likelihood of these risks occurring (Ord, 2020). (iii) to have both cognitive and metacognitive measures (respondents' estimates of other people's judgements) for the same survey items. In completing these aims, we hope to help fill a gap within the literature.

To achieve this, participants took part in a survey in which they were asked to rank a list of ERs (taken from Ord, 2020) based upon the ITN framework and worry. Subsequently, they were asked who they thought was responsible to mitigate the risks and how much they would donate of their income. Following this, they were asked the same questions but what they thought others would put. Finally, they completed a 4-item personality questionnaire (TIPI) (Gosling et al., 2003).

No hypotheses can be made about the ranking of the ERs or the comparison with expert prediction because no previous research has taken place. However, it can be hypothesised that the rankings of participants and what they considered others would rank will correlate due to research based upon egocentric bias (Mildenberger & Tingley, 2019). It can also be hypothesised that a better-than-average-effect (BTAE) will occur when asking participants how much they would donate in comparison with how they think others would (Xiao et al., 2021).

Methodology:

Participants

150 participants were recruited online through opportunistic sampling using Prolific (www.prolitic.com) [Accessed, June 2022]. They were pre-screened and only eligible to participate if fluent in written and spoken English and between the ages of 18 and 90. The sample consisted of 75 women with a mean age of 28.44 (SD=9.69) and 75 men with a mean age of 26.39 (SD=7.56). The locations of the participants are listed in the table below.

Location	Frequency	Percentage
Asia	2	1.3%
Africa	29	19.3%
North America	4	2.7%
South America	1	0.7%
Europe	113	75.3%
Oceania	1	0.7%
Total	150 participants	100%

Table 1: *Frequency of participants' location*

The experiment ran for approximately 10 minutes, and participants were rewarded with a small fee once completing the study (£12.33 per hour). The experiment was conducted using Qualtrics (https://www.qualtrics.com) [accessed June,2022] and participants were given a failed attention check to ensure attention was held throughout the experiment. No data was removed as all participants correctly responded. The study was approved by the ethics committee of the Department of Psychology at the University of York.

Materials:

Qualtrics was used to present the study, and Prolific was used for distribution. Participants were given a brief definition of existential risk (Refer to Appendix A), and the list of existential risks was taken from Toby Ord's (2020) work (Refer to appendix B). To predict participants' political alignment, the items of openness and conscientiousness were taken from the Ten-Item Personality Inventory (TIPI) (Gosling et al., 2003).

Research Design:

A quantitative non-experimental correlational design was used to investigate what the general public considered most worrisome, important, tractable and neglected. These factors were measured by asking participants to rate a list of both anthropogenic and natural existential risks (See Appendix B for list) from most to least for each dimension. (Refer to appendix C for example).

Procedure:

Once participants had opted to take part in the study through Prolific, they were provided with an information sheet and asked to give consent to take part. Once consent was given, participants were asked demographic questions (gender, age and continent location). They were then given the definition of existential risks, proceeding to rank a list of existential risks (Appendix B) on importance, tractability, neglectedness and worry. After ranking the ERs on these separate variables, participants indicated how much of their annual income they would give to help solve (mitigate) their most worrisome risk. Following this, they indicated who they thought had responsibility for reducing existential risks. Subsequently, participants were told to rank the same ERs on what they thought other participants in the experiment would put as most worrisome and predict what annual income others would donate and who they would hold responsible. Participants then completed a 4-item personality questionnaire solely on openness and conscientiousness, taken from the TIPI. To finalise, they completed an attention check and submitted their data.

Results:

Once the data was collected, an average ranking for each existential risk, under each factor, was undertaken. The mean scores for each condition for visual representation are reported in table 1.

	Importance	Tractability	Neglectedness	Worrisome
Climate change	2.53	5.73	4.67	3.14
Nuclear War	3.92	5.73	6.29	3.57
Naturally arising pandemics	4.94	5.93	6.36	5.15
Other environmental damage	5.28	6.47	6.15	6.14
Engineered pandemic	5.73	6.23	6.15	5.47
Unforeseen anthropogenic risks	5.73	6.13	5.75	6.21
Other known anthropogenic risks	5.83	6.43	5.95	5.33
Super volcanic eruption	7.28	5.95	5.86	7.26
Asteroid or comet impact	7.97	5.41	6.41	7.75
Unaligned artificial intelligence	8.19	6.43	6.24	7.79
Stellar explosion	8.96	5.56	6.15	8.18

Table 1. *Mean ranks of each existential risk based on the ITN framework and worry. In this table the lower the number the higher the priority.*

It can be seen from table 1 that Climate change was ranked as the most worrisome, important and neglected. Meanwhile, for tractability, it is Asteroids/comet impact. Another element to mention is the lack of consensus in the rankings for both tractability and neglectedness. This is shown in the table as all ERs are of a similar ranking. This is due to participants ranking them high or low equally often (i.e. lack of consensus).

Kendall's tau-b correlation was then run to determine the relationship between each of the factors, which can be seen in Table 2.

		Worry	Importance	Tractability	Neglectedness
Worry	Correlation Coefficient	-	.855	.019	.112
	Significance (2-tailed)	-	.000	.938	.637
Importance	Correlation Coefficient	.855	-	.019	.075
	Significance (2-tailed)	.000	-	.938	.753
Tractability	Correlation Coefficient	.019	.019	-	-.152
	Significance (2-tailed)	.938	.938	-	.526
Neglectedness	Correlation Coefficient	.112	.075	-.152	-
	Significance (2-tailed)	.637	.753	.526	-

Table 2. *Indicating the results from the Kendall's tau b correlation.*

There was a strong, positive correlation between the ranking of ERs of worry and importance, which was statistically significant ($\tau b = .855$, $p < .001$). However, there was no correlation between any of the other factors ($p > .5$ for all).

Comparison to experts:

In the table below are the overall rankings of the ERs based on worry with Toby Ord's predictions ranked on what is most likely to occur within the next 100 years. Note that importance was not included as it higher correlated with worry's rankings.

The Hazard	Expert rank	Worry rank
Unaligned artificial intelligence	1	10
Engineered pandemics	2/3	5
Unforeseen Anthropogenic risks	2/3	7
Other Anthropogenic risks	4	4
Nuclear War	5/6/7	2
Climate Change	5/6/7	1
Other environmental damage	5/6/7	6
Naturally arising pandemics	8	3
Asteroid or comet impact	9/10	9
Supervolcanic eruption	9/10	8
Stellar explosion	11	11

Table 3: *The ranking of ERs based onToby Ord's predictions and participants' worry. In this table the lower the number the higher the priority.*

A Kendall's tau b was then carried out in order to investigate if the rankings of most worrisome and important correlated with Ord's ranking of the likelihood of existential risks occurring in the next 100 years. The results show there is no correlation between importance or worry ranking when compared to Toby Ord's ($\tau b = .114$, $p = .635$) (both revealed the same results).

The Metacognitive measures:

Worry:

To see if there was a significant correlation between what participants thought, and what they thought others considered most worrisome, a Kendall's tau b test was used. The results indicated a strong positive correlation between personal prediction and prediction of others ($\tau b = .807$, $p < .001$).

Annual income donations:

The annual income portions were averages for both the personal and metacognitive as shown in Figure 1.

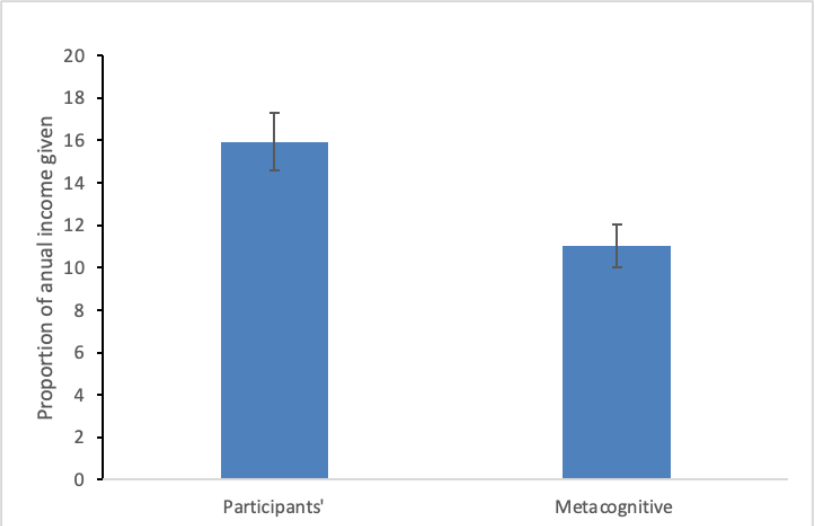

Figure 1*. The average proportion of annual income donated by participants' and what they thought others would respond (metacognitive). The error bars represent +/- 1 standard error.*

The figures suggest that participants believed others would donate less than them. To test if this was significantly different, a paired sample T-test was used. The results indicate that there is a significant difference between both conditions ($t(149) = 4.394$, $p < .001$).

Discussion:

The first aim of the investigation was to gain insight into which risks were seen to be greatest through the ITN framework and worry. The results suggested that climate change, nuclear war and naturally occurring pandemics were the top three most worrisome and important. Participants may have based their rankings on lived experiences. As effects of climate change are becoming more prevalent, threats of nuclear war occur with heightened conflict, and a recent global pandemic killing hundreds of thousands of people, it can be assumed that these risks are at the forefront of people's concerns. When testing to see if there was a correlation between each dimension of the ITN framework and worry, importance was shown to be the driving factor rather than what was most neglected or tractable. It must also be mentioned that for both tractability and neglectedness, participants ranked ERs at both extremities, meaning there was a lack of consensus. Future research should investigate the reasons why participants ranked the ERs the way that they did for each dimension.

When considering the second aim which was comparing the general public rankings with experts' forecasts, the results were concerning. There was no correlation between public rankings and expert forecasts. The most drastic difference was the ranking of unaligned AI. Ord gave it the highest likelihood of occurring in 100 years and the general public ranked it as second to last. These results suggest a lack of awareness from the public about the danger of unaligned AI. Wiener (2016) highlighted the issue of the "tragedy of the uncommons" which involves the misperception and mismanagement of rare ERs in public perception. This leads to neglect of "uncommon risks" which is what can be seen from the results of this investigation. However, a criticism of this study is that the expert comparison was taken from a list of probabilities that the ERs will occur in the next 100 years, rather than a rank from least to most likely to occur. To have a more in-depth analysis and comparison, future research should investigate how experts would respond to this survey.

The third aim was to investigate the metacognitive measures. It was revealed that there was a significant correlation when participants were asked to rank the ERs they considered most worrisome and what they thought participants would rank. This means that the null hypothesis can be rejected as egocentric bias did occur as the rankings were the same. For the second hypothesis, the results showed a significant difference between the proportion of annual income participants would donate and the proportions they thought others would give, in which they thought others would donate significantly less. This refutes the null hypothesis as the better-than-average effect (BTAE) was demonstrated. Previous research (Kim & Han, 2022), has shown that the BTAE was moderated by allocentric goals and negative emotional valence towards others. Future research should aim to understand why some participants think they are more altruistic than others in relation to mitigating ERs and if this may be related to individual differences like culture, age and gender.

In conclusion, the existential risks that respondents most worried about were climate change, nuclear war and naturally occurring pandemics. The degree of worry was strongly related to perceived Importance, but not to Tractability or Neglectedness. There was no correlation between expert forecast and participants' which may be due in part to the "tragedy of the uncommons". The observation that Unaligned AI ranks as the greatest threat in expert opinion, and the least threat (except Stellar Explosions) in public opinion identifies a specific opportunity for improving risk communication. Finally, both metacognitive biases were shown which leads to more questions arising such as; Why do people answer differently and what is it that makes them do so? Future research should test for more individual differences such as political orientation, age, gender and culture to give insight into why participants answered the way that they did. Finally, experts should be asked to rank ERs as it would allow for a more accurate comparison with the general public.

Bibliography:

Bostrom, N. (2002). Existential risks: analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology*, *9*.

Bostrom, N. (2013). Existential Risk Prevention as Global Priority. *Global Policy*, *4*(1), 15–31. https://doi.org/10.1111/1758-5899.12002

Cotton-Barratt, O., & Ord, T. (n.d.). *Existential Risk and Existential Hope: Definitions*.

Gainsburg, I., Pauer, S., Nawal, A., Aloyo, E., Mourrat, J.-C., & Cristia, A. (2021). *How Effective Altruism Can Help Psychologists Maximize Their Impact.*

GOLDBERG, M. H., GUSTAFSON, A., BALLEW, M. T., ROSENTHAL, S. A., & LEISEROWITZ, A. (2021). Identifying the most important predictors of support for climate policy in the United States. *Behavioural Public Policy* , *5*(4), 480–502. https://doi.org/10.1017/BPP.2020.39

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

Kim, M. Y., & Han, K. (2022). For me or for others? The better-than-average effect and negative feelings toward average others during the COVID-19 pandemic. *Current Psychology*, *1*, 1–9. https://doi.org/10.1007/S12144-021-02548-Z/FIGURES/2

Mildenberger, M., & Tingley, D. (2019). Beliefs about Climate Beliefs: The Importance of Second-Order Opinions for Climate Politics. *British Journal of Political Science*, *49*(4), 1279–1307. https://doi.org/10.1017/S0007123417000321

Ord, T. (2020). *The Precipice. Existential Risk and the Future of Humanity*. New York Hachette Books.

Wiener, J. B. (2016). The Tragedy of the Uncommons: On the Politics of Apocalypse. *Global Policy*, *7*, 67–80. https://doi.org/10.1111/1758-5899.12319

Xiao, Y., Wong, K., Cheng, Q., & Yip, P. S. F. (2021). Understanding the Better Than Average Effect on Altruism. *Frontiers in Psychology*, *11*, 3592. https://doi.org/10.3389/FPSYG.2020.562846/BIBTEX

Appendix:

Appendix A:

Definition from the survey about ERs:

“Existential risks are risks that have the potential to permanently destroy human civilisation.” (Bostrom, 2002)

Appendix B:

The list of toby ord’s forecasts (taken from p.162) of “The Precipice. Existential Risk and the Future of Humanity”

Existential catastrophe via	*Chance within next 100 years*
Asteroid or comet impact	~ 1 in 1,000,000
Supervolcanic eruption	~ 1 in 1,000,000
Stellar explosion	~ 1 in 1,000,000,000
Total natural risk	**~ 1 in 10,000**
Nuclear war	~ 1 in 1,000
Climate change	~ 1 in 1,000
Other environmental damage	~ 1 in 1,000
Naturally arising pandemics	~ 1 in 10,000
Engineered pandemics	~ 1 in 30
Unaligned artificial intelligence	~ 1 in 10
Unforeseen anthropogenic risks	~ 1 in 30
Other anthropogenic risks	~ 1 in 50
Total anthropogenic risk	**~ 1 in 6**

Appendix C:

Example of Existential risks ranking layout in the survey