Communication of Existential Risks to Dutch General Public

Existential Risk Observatory

Holly Warner

December 2021

TABLE OF **CONTENTS**

Abstract

A social media campaign strategy was set in place by the Existential Risk Observatory with the aim of communicating existential risks to the Dutch public. This set of campaigns spanned a total of three months, and were primarily conducted over social media (Twitter as the main platform). The outreach strategy was undertaken to drive users to the NGO website to sign a petition calling for AI safety regulation to be incorporated into The Netherlands Dutch Digitalisation Strategy. The specific research question addressed was what type of messaging, utilised to communicate existential risks to the public, best reached users online. Results obtained showed that messaging that was categorised as positive and actionable had the highest engagement rate with users, and those with a theme of climate change and artificial intelligence. We hypothesise that this is due to the public awareness of climate change and artificial intelligence, which refers back to a small qualitative study of twenty users conducted by the Observatory prior to the campaign launch. This campaign has improved our overall understanding of communication pitfalls when translating existential risk issues toward a general audience.

Introduction

An existential risk is "one that threatens the premature extinction of earth-originating intelligent life, or the permanent and drastic destruction of its potential for future development" (Bostrom 2013). Existential risks are often categorised based on whether they emanate from nature or from humans themselves (Häggström, 2016; Bostrom & Cirkovic, 2011). Existential risk can be grouped into natural or anthropogenic risks.

Existential risk is a complex category, the Observatory identifies six groups; artificial intelligence; manmade pandemics; nuclear war; other manmade risks; climate change; and total natural risk. There can be said to be strong public awareness of existential risks such as climate change, and more recently, pandemics with the ongoing COVID19 crisis. Artificial intelligence is also embedded within the public lexicon[1], although this may not include the safety risks of unaligned artificial intelligence and the risks of AGI.[2]

Existential risk requires a proactive approach (Bostrom 2013), and global governance (Ord 2020) in the application of policy to circumvent particular existential and global catastrophic risk scenarios. However, significant development of these mitigation strategies require policy bolstered by public support to generate effective change in attitudes and attention toward alleviation of these risks. Turchin & Denkenberger (2018) highlights the critical role communication within public and policy spheres play in shaping existential risk research.

With respect to the accelerating rate of emerging technologies, it has become apparent that traditional governance is no longer able to keep pace with the rate of development. In order to increase awareness of these risks which aim to translate into actionable policy and regulation

[1] For an analysis of the association between risk and artificial intelligence in the public awareness of artificial intelligence via data gathered from Twitter see Neri and Cozman (2020).
[2] Baum (2018) notes that "it also can be argued that the general public is not the most important group to be informed."

plans to create change, it is a core aim of the Existential Risk Observatory to highlight these risks, distil recent developments in the research of existential risks to easily communicable texts to increase awareness in the public debate. Additionally, an informed general audience may plausibly lead to increased funding for existential risk research, including AI safety research which currently has an asymmetrical budget in comparison to AI development. Increased awareness of existential risks may lead to a diversity of organisations working on risk mitigation.

Over a span of three months the Existential Risk Observatory ran multiple online campaigns to increase awareness of existential risk issues to the Dutch general public. With this communication strategy we aimed to trial different messaging themes communicating existential risks toward a general audience. The aims of this engagement were as follows: to increase existential risk awareness in The Netherlands; to increase awareness of the Existential Risk Observatory itself; to find the most effective messaging strategy to communicate existential risk; and to generate public support for a series of policy amendments to be integrated into the Dutch Digitalisation Strategy in 2022.

Method

An exploratory communication strategy was developed by the Existential Risk Observatory in a series of meetings drawing on a small qualitative study with members of the general public that had no prior exposure to the area of existential risk as it is conceptualised as a field. This study engaged twenty members of various ages and genders in The Netherlands. The primary method consisted of semi-structured interviews with a convenience sample of twenty participants.

The responses obtained from these interviews formed the basis of a strategy to communicate risks to a general audience in The Netherlands. Following the presentation of this qualitative data we formulated a set of twenty messages (to be disseminated via the Observatory's social media accounts) that consisted of a stated risk, a link to the public lexicon, and an actionable statement. These messages had differing thematic bents, and could be attributed to a positive or negative packaging. We determined positive or negative packaging by leading with the statistical probability with no actionable recourse (negative), or by drawing from consensus messaging as utilised by climate change communications by posing the user as capable of acting on this information by doing something. These messages were linked to a petition hosted on the Observatory's website.

Craig's (2018) reflection on risk communication focuses on the role of communication and framing. She notes, "Choice of language is at the heart of the challenge of engaging policy-makers and publics with risk, both mundane and extreme." The petition was hosted via our website calling on the Dutch government to take action on artificial intelligence safety regulation. We utilised Twitter as our main platform for the dissemination of this campaign. The individual tweets were grouped together in monthly runs. These campaigns were boosted with a budget of 25 Euro per day whilst ongoing. We experimented with authority messaging, deploying citations and quotes from authority figures in the field of existential risk and beyond, and also drawing on the Observatory's op-eds in traditional print media.

Results

Iterations of the messaging campaign drew from the first run in October of ten messages total with a link to the petition hosted on the Observatory website. Campaigns were run over a weekly period per tweet launched, and had a budget of 25 Euro per day to garner impressions.

CAMPAIGN SPENDING		
Both bid strategies set to auto bid (recommended) Both frequency cap: Automatically optimise ad frequency (recommended) Daily budget was set at €25 (October) Gender: Any Age: 18 and up Location: Netherlands. Language: Dutch		
OCTOBER	**NOVEMBER**	**DECEMBER**
TOTAL: €180.00	TOTAL: €370.00	TOTAL: €45.00
Targeting features: Artificial intelligence, GCR, Existential risk	Targeting features: Artificial intelligence, GCR, Existential risk	Targeting features: Artificial intelligence, GCR, Existential risk
Interests: Technology + Computing: Computer Programming and Tech News	Interests: Technology + Computing: Computer Programming and Tech News	Interests: Technology + Computing: Computer Programming and Tech News
Audience Estimate: 1.0M - 1.1M	Audience Estimate: 970.7K–1.0M	Audience Estimate: 970.7K–1.0M

Table 01. Campaign spending and attributes by month

The tweets represented in the tables below can be found in full in the appendix. It is worth noting here that the best and worst performing tweet across the campaigns was the same tweet:

1. The future of humanity has incredible potential. However, we can waste all this potential if we're making mistakes now. The actions you can take now are uniquely important. Let's take the best actions together. Sign our petition here:

When combined with the accompanying image (see appendix) the second run of this message received no link clicks. The figures in the table below have been adjusted to approximately 12k impressions to link clicks:

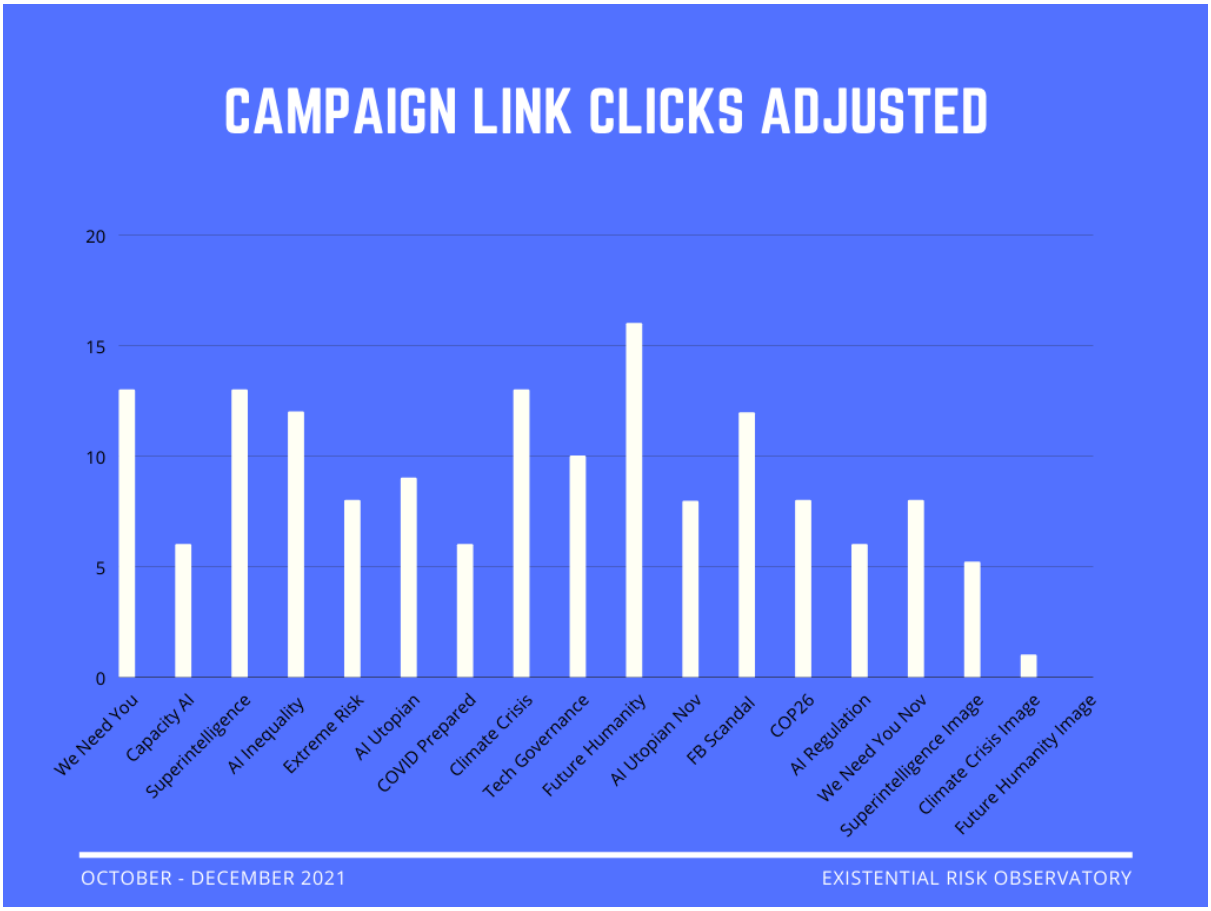

Table 02. Total clicks to petition with adjustments for impression numbers

We adjusted the number of link clicks to approximately 12,000 impressions as average over the eighteen tweets that were optimised toward impressions. We excluded the videos and authority messaging as these were not defined by comparable parameters. Following meetings with the Observatory's social media strategist we experimented with approaches to campaign strategy. We aimed to maximise our conversion rate during the campaign sequence, it was noted that a good conversion rate would yield 2-3% as the standard for a well implemented social campaign.

Based on the key eighteen tweets above we achieved an average follow-through to link click (on the petition) of 8.561 clicks to 12k impressions. When we reran three previously high performing tweets with images this follow through dropped by approximately 85%. It is difficult to assess this significant drop in engagement due to a multitude of factors but we posit that audience exposure to the messages previously had been very close in timing to the rerun. Sum of all total impressions including media and authority messaging but excluding the rerun (Future of Humanity) connected to alternate petition is 315, 081. The average impression count was 12, 118.50 with the median sitting at 12, 722. A total of 212 link clicks span the three campaigns. The most effective message had a follow through of 12.5% more than the average.

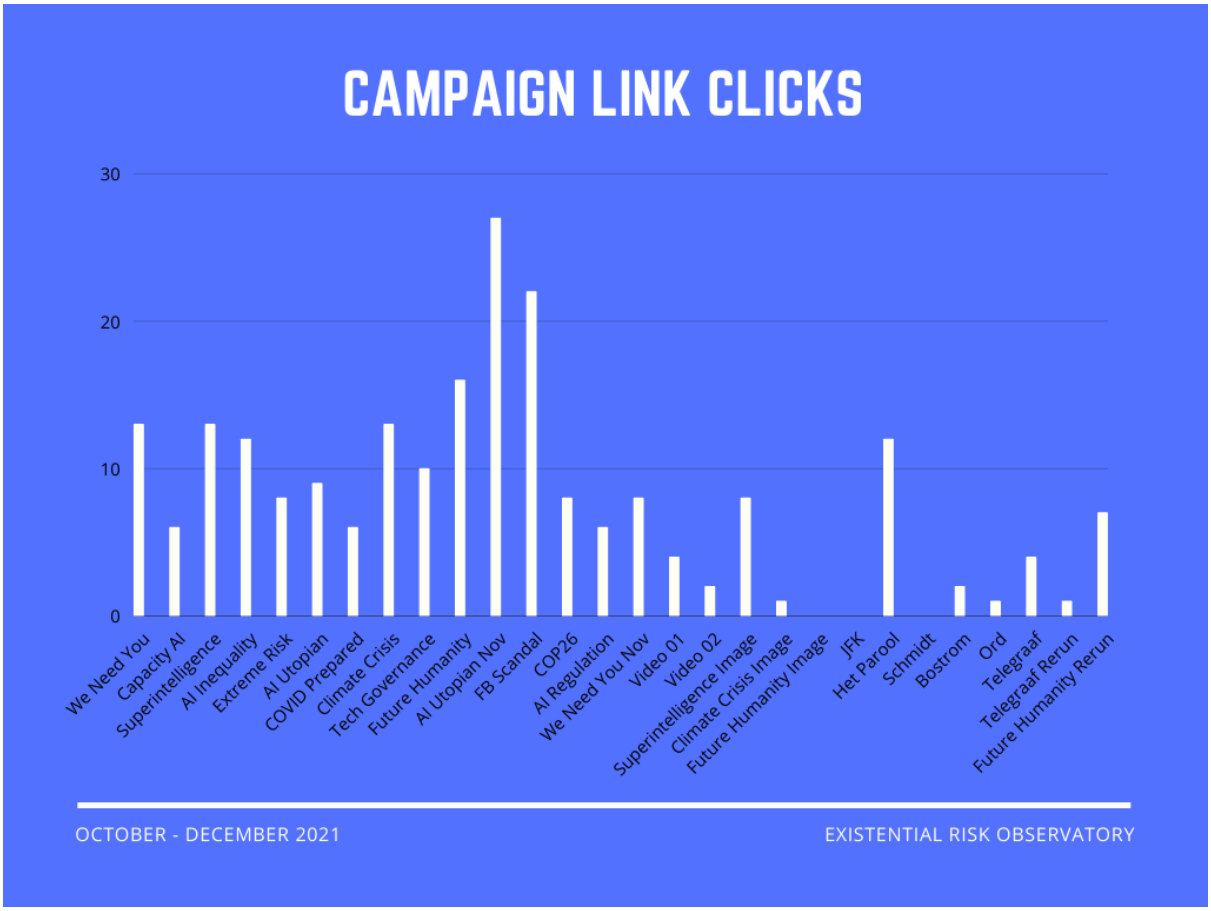

Table 03. Amount of link clicks per tweet

In the October iteration of the campaign the highest performing tweet had 16 link clicks. The petition at the end of this promotion had 7 signatures total. The impressions gained from the campaign were very high, at 116, 951 total but this did not translate to follow through in terms of click through metrics or signatures to the petition. The two top-performing tweets are themed around artificial intelligence and climate change. The top performing tweets were those with a positive message bent.

The first November iteration of the campaign included five messages modelled on the theme of top performing types in the October campaign. We also reran the three top performing tweets of the last iteration, this time including an image designed toward the theme of the tweet. The highest performing tweet had 27 link clicks. Impressions increased during this campaign to 136, 000. The top performing tweets of this campaign were the newer messaging type, with a theme of artificial intelligence, and one in particular that was tied to the current Facebook scandal at the time of publishing. Including five new tweets, three of the top October tweets rerun with images that we made, we also experimented with two short videos optimised for the platform outlining the existential risk of AI and including footage of Elon Musk discussing this.

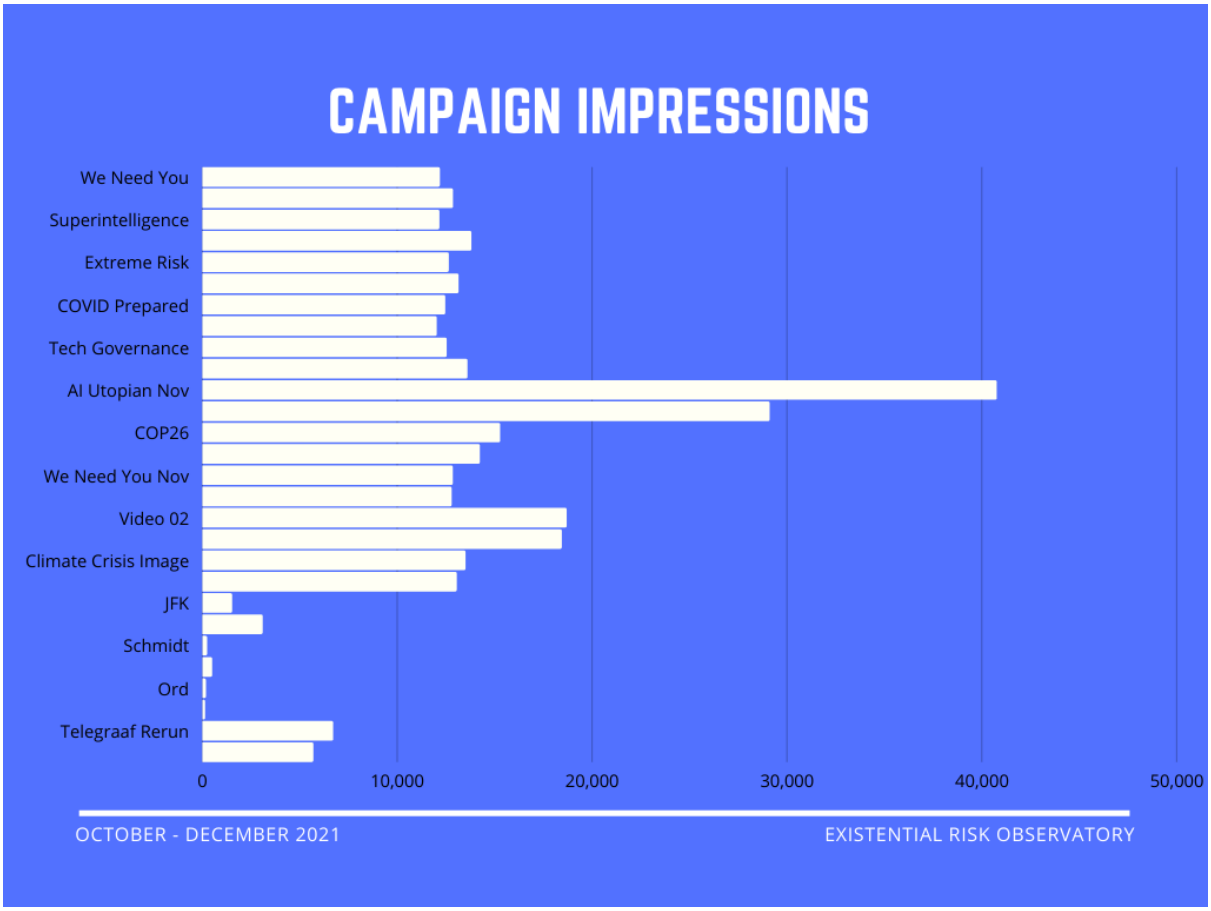

Table 04. Amount of impressions reached per tweet

For the latter November campaign we reoriented our approach based on the results of the previous two campaigns and with input from our social media strategist. For this iteration we trialled a form of 'authority' messaging, solidifying the weight of the messaging on existential risk. We experimented with different formats of authority posts that add weight. The November authority iteration differed in metrics to the previous campaigns. This was optimised for 'Reach' rather than 'Engagements'. This resulted in substantially fewer impressions across the board.

The final run of the winter 2021 social media campaign was centred around rerunning the best performing authority messaging tweet, which was an Observatory opinion article that ran in the print and online output of De Telegraaf. Alongside this op-ed post we also boosted our overall best performing message again, which was a positive message on the future of humanity. This message was accompanied by the image added in for the November iteration, and linked to a petition that we hosted on Avaaz, as opposed to our own website.

Discussion

As illustrated above, the follow through from impression count to link click was not very successful. We followed the campaigns with reflection and tailored our approach. Following a consultation with our social media strategist we aimed for a smaller audience that is really qualified to begin with. Going forward we are trialling methods to target key members of this audience. We have formulated a set of strategies going forward: run a Facebook campaign to compare results; experiment with different ad groups for climate change and AI messaging; and work on defining our audience for Facebook in particular.

Following team social strategy meetings we decided to target similar follower profiles, such as Future of Life. We tried different ad groups and experimented with images/animation or text only to test results. We experimented with different audiences and methods and kept track of the outcomes to devise the best approach. Experimental difficulties include the petition being hosted on our website. In meetings with our social media consultant, he agreed that this may be seen as trying to parse signatories for data or may be potentially interpreted as an online scam. Other perceived difficulties include the relative weight of the organisation in terms of authority, as compared to established research institutes such as CSER or FHI, and that the campaigns were launched in quick succession, potentially being promoted to the same audience repeatedly, with the same parameters so closely in tandem. Additionally we noted that the topic of existential risk is complex to communicate, this is compounded by a multitude of factors. The Observatory's Dutch Twitter account is new and lacks an extensive network; the ongoing COVID pandemic has shifted audience focus significantly; and existential risk may be too negative or unknown a topic to translate to a general audience through the character limit on Twitter.

Conclusion

Existential risk communication to the public required the organisation to work without precedent. Additionally the Observatory is a new organisation with no prior reputation to draw from when reaching out to the Dutch general public. We trialled different groupings, techniques, and campaigns with the goal of eliciting data that would allow us to move forward with a robust campaign strategy for future communications. The messaging that had the highest conversion rate to impression count was thematically located in artificial intelligence and climate change, two areas that are already firmly embedded in the public awareness. We posit that this gave us more of an ability to connect with our audience on these issues, rather than other existential risks such as emerging technologies such as bio and nanotechnology. We had solid interaction with our messaging that highlighted our content output in print media such as Het Parool and De Telegraaf. The addition of this outside certification of the organisation may have helped to reassure our target audience of the Observatory's legitimacy. Tentative conclusions are that the narrative of messaging shared on social media with relation to existential risk does matter, but not to a significant extent, and particular themes may work better than others, as discussed above. It is also worth highlighting that messaging that is linked into an issue present in the current news cycle, such as the Facebook scandal, appears to engage users more. These are preliminary conclusions as the scope of our social media campaign was limited. Future campaign iterations with well established parameters would increase the robustness of outcome results.

References

Baum, S. D. (2018) "Countering Superintelligence Misinformation." Information. An International Interdisciplinary Journal 9 (10): 244.

Bostrom, N., 2013. Existential risk prevention as global priority. *Global Policy*, *4*(1), pp.15-31.

Bostrom, N, and M. M. Cirkovic. 2011. Global Catastrophic Risks. Oxford: OUP.

Craig, C. (2018). Risk management in a policy environment: The particular challenges associated with extreme risks. Futures.

Häggström, O. 2016. Here Be Dragons: Science, Technology and the Future of Humanity. Oxford: OUP.

Neri, H. and Cozman, F., 2020. The role of experts in the public perception of risk of artificial intelligence. *AI & SOCIETY*, *35*(3), pp.663-673.

Ord, T., 2020. *The precipice: existential risk and the future of humanity* Hachette Books.

Turchin, A. and Denkenberger, D., 2018. Global catastrophic and existential risks communication scale. *Futures*, *102*, pp.27-38.

Appendix (Messages in chronological order)

- We've all seen the polarisation on social media seep into our daily lives, AI could deepen this division in our communities. Act now to ensure our leaders put measures in place to combat this. Sign our petition here:
- From self-driving cars to caring robots, the potential problem-solving capacity of AI is as exciting as it is mind-boggling. But as we develop the capabilities of intelligent machines, we must address the potential dangers that lie in front of us. Sign our petition here:
- When AI develops into superintelligence, power will be concentrated in the hands of a few. How stable a world would this create? Let's make our government check AI development to ensure that doesn't happen. Sign our petition here:
- AI is currently leading to more inequality. We can expect AI to become a lot more capable, fast. We need our governments to check AI development to make sure AI will benefit us all. Sign our petition here:
- Now is the time to give extreme risk our full attention, we need our leaders to act immediately to protect the future of humanity. Sign our petition here:
- AI could be used to create a utopian future for the benefit of all, but if left unregulated it may produce a dystopian future for humanity. You can change this, sign our petition here:
- We were not prepared sufficiently for climate change and COVID, let's not do the same thing with artificial superintelligence. Let's make sure we decrease the next existential risk as much as possible. Sign our petition here:
- The climate crisis has proven how everyday people can make big changes. So let's push that same energy and activity toward decreasing the biggest threats to humanity. Sign our petition here:
- The unprecedented rate of development of new technologies, such as AI, far outpaces the speed of adaptation of traditional governance. We must pivot to tackle these problems head on. Sign our petition here:
- The future of humanity has incredible potential. However, we can waste all this potential if we're making mistakes now. The actions you can take now are uniquely important. Let's take the best actions together. Sign our petition here:
- AI could be used to create a utopian future for the benefit of all, but if left unregulated it may produce a dystopian future for humanity. You can change this, sign our petition here: (November)
- The Facebook scandal has shown us how biased algorithms polarise our society. Let's work together to make sure AI is regulated for our safety. Sign our petition here:
- The COP26 summit is underway with climate negotiations, the action we take now can stop the biggest threats to humanity. Sign our petition here:
- Instead of us being used by tech, we should use tech to bolster our future. Sign our petition here to call on our government for regulation on AI:

- We've all seen the polarisation on social media seep into our daily lives, AI could deepen this division in our communities. Act now to ensure our leaders put measures in place to combat this. Sign our petition here:
- When AI develops into superintelligence, power will be concentrated in the hands of a few. How stable a world would this create? Let's make our government check AI development to ensure that doesn't happen. Sign our petition here: (See Image.02)
- The climate crisis has proven how everyday people can make big changes. So let's push that same energy and activity toward decreasing the biggest threats to humanity. Sign our petition here: (See Image.01)
- The future of humanity has incredible potential. However, we can waste all this potential if we're making mistakes now. The actions you can take now are uniquely important. Let's take the best actions together. Sign our petition here: (See Image.03)
- Zoals JFK opmerkte: "Onze vooruitgang in het gebruik van wetenschap is groot geweest, maar onze vooruitgang in het ordenen van onze relaties klein". Laten we samenwerken om existentiële risico's te verminderen. Teken hier onze petitie:
- Kunstmatige intelligentie maakt een almaar groter deel uit van ons leven. En dat is niet zonder risico's, betogen wij in Het Parool. Vind jij ook dat de overheid AI existentieel risico in de gaten moet houden? Teken dan nu onze petitie:
- "Reductie van existentiële risico's is één van de belangrijkste uitdagingen waar de mensheid voor staat", zegt Andreas Schmidt, universitair docent bij de Rijksuniversiteit Groningen. Laat de overheid deze risico's reduceren! Teken nu onze petitie:
- Professor Nick Bostrom zegt dat 'zelfs mensen die praten over het broeikaseffect nooit de bedreiging door AI noemen'. Teken hier om deze bedreiging te verminderen:
- Toby Ord van de Universiteit van Oxford schat dat we een kans van 1/6 hebben op het uitsterven van de mens in de volgende eeuw. Maar de overheid kan veel doen om onze kansen te vergroten! Teken hier onze petitie:
- Kunstmatige intelligentie streeft onze hersenen over enkele jaren voorbij in complexiteit, stellen wij vast in De Telegraaf. De risico's die dit met zich meebrengt zijn onacceptabel. Teken dan hier onze petitie:
- Kunstmatige intelligentie streeft onze hersenen over enkele jaren voorbij in complexiteit, stellen wij vast in De Telegraaf. De risico's die dit met zich meebrengt zijn onacceptabel. Teken dan hier onze petitie: (Repeated)
- The future of humanity has incredible potential. However, we can waste all this potential if we're making mistakes now. The actions you can take now are uniquely important. Let's take the best actions together. Sign our petition here: (Repeated)

Appendix (Images attached to messages)

Image 01. 'The climate crisis'

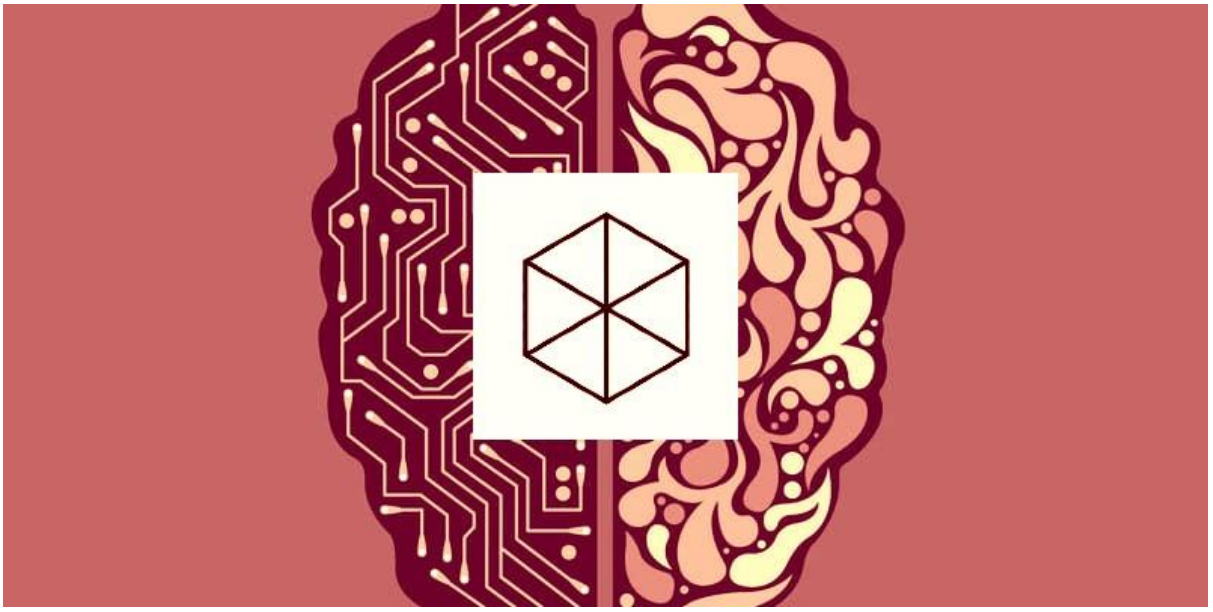

Image 02. 'When AI develops into superintelligence'

Image 03. 'The future of humanity has incredible potential'