Avaaz response to the White Paper on Artificial Intelligence - A European Approach

Introduction

"We believe Europe needs to step up in order to balance both protecting Democracy and protecting Freedom of Speech, because if we don't, at some point disinformation will regulate us." Christoph Schott, Avaaz Campaigns Director[1]

Avaaz is the world's largest online civic movement. Our 62 million members campaign for urgent action on the key issues of our time - the climate crisis, ecological collapse and the erosion of democracy. Overcoming these threats will, above all, require wisdom from decision makers.

AI has a positive role to play in all these issues and many more. Its potential to transform society surpasses the agricultural, industrial and internet revolutions because, unlike these previous technologies, it introduced a powerful new decision maker alongside humans for the first time- the AI automated decision.

Avaaz has been investigating and campaigning on the threat posed[2] by disinformation, organized and distributed at scale by AI on social media platforms. Commissioner Jourova's "*call to arms*"[3] to combat disinformation earlier this year, recognised the seriousness of the risks to democracy. She called again for action on June 10 saying "*To fight disinformation, we need to mobilise all relevant players from online platforms to public authorities, and support independent fact checkers and media.*"[4] To this list of actors who can make a difference in the fight against disinformation we would add the AI Framework, it's architects cannot miss out on the opportunity to understand, and regulate the role of AI in disseminating disinformation.

We encourage the Commission to take the lead now in setting standards that not only minimise risk but create the conditions in which AI can contribute to, rather than undermine, the wisdom humanity will need to wield in the coming decades for all life to flourish and thrive on Earth. The Commission's power is to connect the relevant pieces of the digital future framework to protect the multiple freedoms that disinformation attacks.

[1] Interview Euronews 10 June in response to the EU's decision to strengthen action to tackle disinformation https://www.youtube.com/watch?v=xYCxQ9TXFQc

[2] We have included a summary of our extensive research in Appendix 1

[3] See https://ec.europa.eu/commission/presscorner/detail/en/SPEECH_20_160

[4] https://ec.europa.eu/commission/presscorner/detail/en/ip_20_1006

Contents of this Submission

We have provided answers and substance in our response to the Commission's questionnaire. This paper addresses 2 key issues we were unable to cover in the questionnaire: - (i) inclusion of AI use in the dissemination, recommendation and selection of media content as a high risk for regulation under the AI Framework and (ii) the need to regulate AI use to achieve accountability and transparency with a focus on the spread, at scale, of disinformation.

We have also included three Appendices to support our recommendations further:

Appendix 1: Current Disinformation threats and Avaaz reports summaries - including substantive evidence on the how AI and platform algorithms spread it;

Appendix 2: The rights, freedoms and values impacted by AI used in content distribution and curation that require it to be treated as a high risk activity; and

Appendix 3: An regulatory audit framework for AI to combat disinformation

AI used in the dissemination, recommendation or selection of content at scale to users is a high risk processing activity that should be regulated in the AI Framework.

"We are racing towards a world in which we are allowing "the machine to know us better than we know ourselves*, and we may no longer have the capacity to know whether or not we have something to hide, from whom or how to hide it."*
---Susie Alegre, international human rights barrister and Associate - Doughty Street Chambers

In our view the proposed AI framework will not achieve a future proof position if it adopts a view that regulation can be based on a bullet list of perceived risky AI applications. The framework should be developed to protect against unintended but demonstrable risks/harms due to AI, irrespective of initial risk levels in a certain sector. If however the Commission retains this approach then AI use in the dissemination, recommendation and selection of media content must be included as a high risk application. This is due not the least in its role in the spread of disinformation on social media and other content recommendation platforms.

The power of disinformation, as Commissioner Jourova warned, "*to blur the lines, to polarise and make us indifferent*" is massively amplified by the information processing capacity of AI in the dissemination, recommendation and selection of media content. This capacity has the power to undermine fundamental rights like freedom of thought, from which flows freedom of opinion, expression and information, robbing humanity of its wisdom.

There is growing understanding, from scientific, legal and civil society as to the extent to which our freedom of thought, our "forum internum", is being impacted by the opaque decision-making of AI.[5] The promotion of content through AI recommendation algorithms is invisible to the service user. The opaque AI decision-making that creates your news feed or recommends content provide any rationale to the service user as to why they have been served a particular piece of content by a social media platform's AI. If the way AI delivers content to us - through non transparent collection and analysis of our personal data and the economic valuation of attention above all else, means we lose the right to the full picture, the unadulterated diversity of human experience, then we are not truly free to think. If we are not free to think, then all the other rights - such as freedom of speech, opinion and information, that freedom of thought provides a foundation for, cannot be realised never mind adequately protected.

[5] E.g Susie Alegre (ibid); Amnesty International "Surveillance Giants" 2019 Report (https://www.amnesty.org/en/documents/pol30/1404/2019/en/); Chatham House "Online Disinformation and Political Discourse: Applying a Human Rights Approach" and (https://www.chathamhouse.org/publication/online-disinformation-and-political-discourse-applying-human-rights-framework); and BBC Radio 4 "Forum Internum" (https://www.bbc.co.uk/programmes/m000fq3y)

We have included a detailed description of how our freedom of thought is impacted by social media's AI in Appendix 2. Drawing on examples such as Cambridge Analytica, the tragic suicide of Molly Russell in the UK, and Facebook's sale of data about Australian teenagers mental states, we outline how AI - designed with the economic aim of maximising the attention of the user, and time spent within the platform's exclusive environment is polarising us into silos and undermining our judgement. It is worth considering this impact now, as AI regulation may be the only credible long-term option to regulate such harms at scale, particularly when end to end encryption is added to the mix. We know that some social media platforms for example are pursuing encryption at pace, even though it will make their current efforts to combat disinformation, nigh on impossible. Just last week Facebook's representative Monica Bickert admitted to the UK Parliament that Facebook are pressing ahead with their end to end encryption plans, even though they have no current way to secure the protection of children from online exploitation through the platform once encryption kicks in.[6]

In terms of equality, Appendix 2 includes detail on how the AI that drives the bulk of social media's moderation is designed with wholly inadequate data sets to enable it to do the job. Automated moderation uses confirmation machine learning models that need a minimum threshold of classified data (data sets) in order to accurately identify and flag illegal content or content that breaches their terms and conditions. However, we found clear examples in Facebook's Indian platform that excluded linguistic and ethnic minorities from being protected against hate speech targeting them. We found for example that the language models of Facebook in India did not extend to the Assamese language, and therefore could not recognise common hate speech terms used against the Bengali Muslim minority in Assam. As the data set was inadequate to build an Assamese classifier, the AI deployed to detect hate speech could not cover hate speech in the Assamese language, and so it spread without check. Since we reported this to Facebook they have now built a glossary of 50,000 Assamese words, but we would question how many other minorities are absent from the AI's data sets.

In this way AI driven decision making could, over time, increase the asymmetry of power between human users of technology and its developers, human or otherwise. And if this all seems a distant risk, far off, we would point to the rise of disinformation across AI driven platforms and ask that the EU adopts a 'precautionary' regulatory approach in order to avoid or reduce harms that are plausible. It should do so here by recognising the impact of this asymmetry on the most fundamental of freedoms. Commissioner Jourva has said that the EU's actions *" are strongly embedded in fundamental rights, in particular freedom of expression and information."* Recognising that these flow from freedom of thought, designing the AI framework to protect these rights is the only way for it to keep pace as the threat evolves.

We also address in Appendix 2 how data rights and product liability are impacted by AI and note in comparison to the EU framework that the UK regulatory body, the ICO is currently producing guidance on how the design and operation of AI affects data use.

[6] See https://www.parliamentlive.tv/Event/Index/4f63eb5e-834a-403d-bbeb-f40d903ef173 at 16.06pm

We applaud Commissioner Jourova's recent statement that "*the role of public authorities is not to interfere with content policies of private companies but to ensure that fundamental rights are protected online as well as offline — rights such as freedom of expression and information, non-discrimination, right to security*".[7] We believe such an approach must include and regulate the AI involved dissemination, recommendation or selection of content to users at scale as a high risk under the AI Framework set out in the Commission's White Paper.

Why must the future EU regulatory framework for AI regulate the reach of disinformation?

*"In today's technology-driven world, where warriors wield keyboards rather than swords and targeted influence operations and disinformation campaigns are a recognised weapon of state and non-state actors, the European Union is increasing its activities and capacities in this fight." "While online platforms have taken positive steps during the pandemic, they need to step up their efforts".*Vice-President for Values and Transparency Věra Jourová 10 June 2020

Whilst we agree with these sentiments, the infodemic curve won't be flattened just by asking social media platforms to be more transparent. At a time when lies about Covid-19 are literally costing lives, Brussels cannot keep shying away from the heart of the debate: Regulate disinformation, or one day it will regulate us. The second theme of this submission looks deeper into what integrated and systemic regulatory solutions for AI would look like to deal with Disinformation, and why it is essential these are featured in the AI Framework. Transparency and accountability needs to be mandatory, not suggested as voluntary options or compartmentalised into, for example, the Disinformation Code of Practice or the Democracy Action Plan.

These issues need urgent, integrated and systemic policy solutions, a narrow view of risk minimisation simply won't do.

How does AI play a role in Disinformation online?

Platforms use AI in the form of algorithms to determine when and in what order users see content that the algorithm determines may be of interest to them.[8] Algorithmic content curation like this has important consequences for how individuals find news, but AI takes into account many factors other than the user's own selection of content - for example popularity and the degree to which information provokes emotions like outrage or provides confirmation bias are increasingly important in driving algorithms' choices of which content to promote.[9]

[7] https://www.politico.eu/article/european-commission-vp-backs-twitter-in-trump-battle/
[8] https://medium.com/@guillaumechaslot/how-algorithms-can-learn-to-discredit-the-media-d1360157c4fa
[9] https://kf-site-production.s3.amazonaws.com/media_elements/files/000/000/142/original/Topos_KF_White-Paper_Howard_V1_ado.pdf

This ability to attract attention, without the user's conscious participation, is fundamental to the platform's business model. A New York Times article reported that in 2018 an internal meeting at Facebook outlined its understanding of how it's AI could operate to exploit the attention seeking aspects of our nature. A slide in a presentation at the meeting stated "Our algorithms exploit the human brain's attraction to divisiveness," read a slide from a 2018 presentation. "If left unchecked," it warned, Facebook would feed users "more and more divisive content in an effort to gain user attention & increase time on the platform."[10]

And so the speed and scale at which content "goes viral" is enhanced by AI, it grows exponentially, regardless of whether or not the information it contains is true or promotes harm or hatred. In this way, although the Internet has provided more opportunities to access information, algorithms have made it harder for individuals to find information from critical or diverse viewpoints.[11] Therefore, there is a risk that users get trapped in an an online bubble of disinformation[12], hate speech or suicidal ideation.

We cannot know exactly, given the black box model of digital content recommendation, exactly what values the algorithms prioritise when they make content recommendations. This has made tackling the problem of the spread of misinformation across platforms harder to predict, track and address.

We do know that the platform's algorithms accelerate the spread of viral disinformation equally, whether it's about fake cancer cures, climate misinformation or claims about government policies on immigration[13]. Avaaz has shown how political disinformation and polarisation is pushed by the very same people that push an anti vaxx message. The same problem that risks overturning the global effort that will be needed to end the health crisis also continues to allow fake bots and foreign interference with elections.

A future regulatory framework must connect the dots and provide accountability and transparency over the nature and impact of AI from its design and operation to its output. The Commission must avoid the compartmentalisation of AI policy, pushing the impact of disinformation into other policy silos. As we have argued throughout this response, the AI framework should integrate disinformation, and accordingly the AI employed in the dissemination, recommendation or selection of content at scale to users, into its high risk

[10] https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499?mod=hp_lead_pos5

[11] https://medium.com/trust-media-and-democracy/three-reasons-junk-news-spreads-so-quickly-across-social-media-385b91c8d779

[12] https://www.reuters.com/article/us-alphabet-youtube-content/youtube-sharpens-how-it-recommends-videos-despite-fears-of-isolating-users-idUSKBN1DT0LL

[13] See our report Why is YouTube Broadcasting
Climate Misinformation to Millions? In Appendix 1 and at
https://secure.avaaz.org/campaign/en/youtube_climate_misinformation/

framework - and ensure that framework is kept open to new forms of harm that may emerge over the framework's lifespan.

How can the AI Framework regulate the spread of disinformation?

We note in the consultation that the High-Level Expert Group will be further developing aspects of transparency and human oversight for practical use by companies to be finalised by June 2020. In anticipation of this work, and to ensure it is sufficiently comprehensive, we have set out a summary of the factors we believe will be key to the AI Frameworks success in regulating the spread of disinformation. The continued protection from liability offered in the e-commerce rules should carry these obligations of accountability and transparency[14]

Transparency and Audit

In the context of combatting disinformation, **regulators need information gathering powers to allow independent audits** of the platform's content delivery and monetisation systems and to assess the technical and resource investment of each platform to combat disinformation. We have included a comprehensive audit plan at Appendix 2 to which covers both design and operation of AI and provides the means by which, even with end to end encryption, platforms will have to give transparency on the following:

- the ability of the platform's algorithm to detect and slow down the spread of disinformation using the detox principles outlined above.
- the ability of the platform's AI to do this across the language models of all likely users.
- The ability of the AI's design to facilitate the exercise of individual rights to avoid biased/inaccurate sources of information - for example by recognising and labelling categories of state sponsored media, known conspiracy domains, etc.
- **Provide clarity on the platform's actions in response to issues detected by its AI, including the reasons for any action taken.**
- the ability of the platform's algorithm to detect and deal with breaches of other platform standards which could impact on individual rights;
- the degree to which users data has been processed during any operation which could impact on individual rights, for example to make content recommendations which direct misinformation content to the user, and evidence of the transparency of such processing;
- the ability of the algorithm to facilitate the exercise of individual rights in all automated decision making - such as in the recommendation and search services of social media platforms or the curation of individuals' newsfeeds. Can users turn on or off defined areas of their own data - can users exclude irrelevant data brought into their data processing - such as the content choices of those in their location, income bracket or other demographic category. Is the automated decision making sufficiently explained before the user engages with the service? Has the AI been designed so that the user

[14] To be addressed in the forthcoming Digital Services Act consultation

can still benefit, albeit in a perhaps more limited way from the service if they decide to limit the degree of automated processing of their data?

Accountability

If the information gathering powers outlined above, and detailed in Appendix 3 are included inIn the context of disinformation, accountability covers a spread of behaviours that the framework should enshrine so that the platforms act to halt the spread of disinformation. The AI Framework should require platforms to :

- Improve and invest in the AI that moderates their systems, ensuring it has **adequate data sets and the predictive capacity to recognise hate speech and verified disinformation.** Without such capacity it can neither remove illegal hate speech nor correct the record on disinformation efficiently - beating the viral speed with which it otherwise spreads.[15] Platforms must review these data sets on an ongoing basis, working with reputable fact checking organisations to confirm disinformation detected.

  These steps are integral to comprehensively detoxifying the platforms' algorithms. **"Detoxing the Algorithm" in this way allows the transparent adjustment of social media platforms' content curation algorithms to ensure that they downgrade disinformation, pages belonging to disinformation accelerators and malicious actors, and other harmful content out of their recommendations to viewers.** This would ensure that the recommendation, search, and newsfeed algorithms are not abused by malicious actors, and that disinformation is sidelined rather than boosted by these platforms.

- Remove all economic incentive for disinformation, both from their own advertising revenue streams and for the bad actors themselves - **with demotion of bad actors and their content from recommendation and search engines** and the **demonetisation** of content from consistent bad actors.

- The objective of trustworthy, ethical and human-centric AI can only be achieved by ensuring an appropriate involvement by accountable human beings in relation to high-risk AI applications

In Summary

[15] Avaaz's research has shown that if social media platforms correct the record when their users are exposed to disinformation, it cuts their belief in those lies in half. By working with independent third-party fact-checkers, tweaking algorithms to downgrade disinformation and alerting users social media can lead the way in such media literacy initiatives.

The design of EU regulation will set the tone and parameters for regulation for years to come for automated decision-making capabilities that will wield vast powers over humanity. We have seen the benefit of a connected, consistent regulatory framework with powers to gather information and enforce against breaches of standards in industries as varied as steel production[16] and media content regulation[17]. Those industries have flourished and innovated under comprehensive regulatory systems that anticipate the harms to workers, users and the European Economy. We believe that the Commission's power is precisely to **connect all the relevant pieces of the digital future framework** to provide this connected framework. That means aligning the Commission’s work streams (like the Democracy Action plan, the Digital Services Act, and future framework Artificial Intelligence) so all play a part in **correcting the systemic faults** that allow disinformation to flourish.

With the inclusion of content disseminating and recommendation AI into the high risk framework, and audits to ensure accountability and transparency, this new framework may be our best hope to ensure that AI brings the benefits you anticipate, without allowing the harms to run unchecked.

[16] “The right regulatory framework - EU legislation is essential for the sustainable development and proper functioning of the internal market, for investor certainty and predictability, and for providing a level playing field for the steel industry” European Commission review of steel industry regulation. See https://ec.europa.eu/growth/sectors/raw-materials/industries/metals/steel_en

[17] “The EU's Audiovisual Media Services Directive (AVMSD) governs EU-wide coordination of national legislation on all audiovisual media, both traditional TV broadcasts and on-demand services” see https://ec.europa.eu/digital-single-market/en/policies/audiovisual-media-services

Appendix 1

Current Disinformation threats and Avaaz reports summaries

Disinformation poses an existential threat to human societies. Concerted campaigns to mislead people have the potential to change public opinion, amplify an issue and change the course of elections. The AI intended to recommend and guide users to content that will hold their attention is being gamed through inadequate data sets to push this disinformation out to social media users. Unjust and inequitable policies regarding immigration and climate could emerge out of such lies, damaging the shared understanding of facts that we need for healthy societies. In the age of a pandemic we are now seeing all too clearly that disinformation is also downright dangerous to human life.

Avaaz's research has shown that online disinformation networks are coordinated, deploy fake accounts and mislead people with content that aggravates existing fault lines in respective countries, such as issues around - race, religion, immigration, minorities, caste, gender, sexual orientation, climate change and so on.

So far we have investigated the harms posed by disinformation in the fields of elections, ethnic tensions and hate speech, climate change, and health disinformation. Our reports in summary, with links to the full reports are here:

Political and Electoral Disinformation

1) **YELLOW VESTS FLOODED BY FAKE NEWS OVER 100M VIEWS OF DISINFORMATION ON FACEBOOK** 15/03/2019 see
   https://avaazimages.avaaz.org/Report%20Yellow%20Vests%20FINAL.pdf
   Avaaz called on Facebook to Correct the Record ahead of EU Elections -- with an in-depth study showing how **fake news surrounding the Yellow Vests reached over 100 million views, and how Russia fueled the divide.**

2) **WHATSAPP - SOCIAL MEDIA'S DARK WEB** 26/04/2019 see
   https://avaazimages.avaaz.org/Avaaz_SpanishWhatsApp_FINAL.pdf
   Avaaz continued to ring the alarm bell ahead of the European Elections on the deluge of fake news and hateful memes on WhatsApp, with a crowdsourced effort detecting hundreds of pieces of potential disinformation and a representative survey showing that about **9.6 million Spaniards received such content.**

3) **FAR RIGHT NETWORKS OF DECEPTION** 22/05/2019 see
   https://avaazimages.avaaz.org/EU%20Disinfo%20Report.pdf

Immediately ahead of the European parliamentary elections Avaaz reported to Facebook a total of nearly 700 suspect pages and groups,followed by more than 35 million people and generating over 76 million “interactions” (comments, likes, shares) between January and April 2019. Facebook took down 132 of the pages and groups reported, accounting for almost 30% of all interactions across these networks, and 230 suspicious profiles. **Together, the pages taken down had reached 762 million estimated views over the three months ahead of the elections.**

4) **MEGAPHONE FOR HATE , DISINFORMATION AND HATE SPEECH ON FACEBOOK DURING ASSAM’S CITIZENSHIP COUNT** October 2019 see https://avaazpress.s3.amazonaws.com/FINAL-Facebook%20in%20Assam_Megaphone%20for%20hate%20-%20Compressed%20(1).pdf
This report investigated the extraordinary chorus of abuse and hate in Assam against Bengalis, Muslims intended to influence the political approach to the National Citizenship Count on Facebook. This is the first report that dissects the nature of this online hatred in Assam and warned that such dangerous prejudice must not be allowed to influence policies to strip away citizenship rights from 1.9 million people. This report was our first to expose the limitations of Facebook’s artificial intelligence (AI) driven strategy to detect hate speech.

5) **US2020: ANOTHER FACEBOOK DISINFORMATION ELECTION?** 05/11/2019 see https://secure.avaaz.org/campaign/en/disinfo_report_us_2020/
One year out from the US 2020 elections, this Avaaz investigation uncovered political “fake news” flooding US citizens on Facebook. Politically relevant disinformation was found to have reached over 158 million estimated views, enough to reach every reported registered voter in the US at least once. Over a ten-month period, between January 1 and October 31, 2019 our team analyzed the 100 top fake news stories about US politics fact-checked and debunked by reputable US fact-checking organizations. Collectively, they were posted over 2.3 million times.

Health Disinformation

Is Fake News Making Us Sick? How misinformation may be reducing vaccination rates in Brazil.
November 2019 See
https://avaazimages.avaaz.org/Avaaz%20-%20Is%20Fake%20News%20Making%20Us%20Sick%3F.pdf
In this joint report by Avaaz and the Brazilian Society of Immunization (SBIm) provided new and revelaining data on Brazil’s low take up of immunisation. 13% of the Brazillian’s we polled did not believe in the benefits of childhood vaccination programmes. The majority of these people believed false information on the risks. The report uncovers the networks behind the distribution of anti vaccination misinformation.

How Facebook can Flatten the Curve of the Coronavirus Infodemic
April 15 2020 See
https://secure.avaaz.org/campaign/en/facebook_coronavirus_misinformation/
An international report covering health disinformation in Europe and the US, including vital new data on the networks and spread of coronavirus.

Climate Disinformation

Why is YouTube Broadcasting Climate Misinformation to Millions?
01/16/2020 see https://secure.avaaz.org/campaign/en/youtube_climate_misinformation/

This report revealed how YouTubedrives climate misinformation videos through it's recommendation algorithm - which gives these videos free promotion showing misinformation to millions who wouldn't have been exposed to it otherwise. This report delved into the design of the social platforms algorithms that serve up the faked content through hidden choices about users preferences. It also describes the process of monetisation, that allows fake news suppliers to profit from the videos whilst brands unwittingly pay to show ads alongside them.

Appendix 2

The rights, freedoms and values impacted by AI used in the dissemination, recommendation and selection of media content that require it to be treated as a high risk activity

Freedom of Thought

When we think about how information is sifted, parsed and restricted through the automated decision making of AI, a new human rights paradigm from the consumer rights or data rights emerges, that of the freedom of thought. Freedom of Thought is an absolute right enshrined in the EU Charter of Fundamental Rights and Freedoms, the European Convention of Human Rights, the International Covenant of Civil and Political Rights, the Universal Declaration of Human Rights and the American Convention on Human Rights, which protects individuals from incursions on their innermost sanctum -- their forum internum -- and is the progenitor of many rights. For if we can't guard our own thoughts, how can we exercise our right to freely express ourselves or to speak?

The rights to freedom of thought and the closely related rights to freedom of opinion and information was inscribed into human rights law following World War II, when the drafters of the initial international human rights corpus had a fresh memory of the role that large scale propaganda played in perpetuating the horrors of Nazi Germany. What is different about new technological developments, however, is the *manner* in which they facilitate the amplification of propaganda and microtarget of users. So the risk and potential harm here is the **scale and speed** at which disinformation reaches us, the manner in which the platforms facilitate it, and how the disinformation is tailored to influence each one of us individually. The EU Charter has developed the rights contained in earlier instruments to reflect the evolution of challenges and understanding of particular rights and the right to mental integrity included in the Charter can be viewed as an additional aspect of the right to freedom of thought in the modern context.

How are these rights engaged?

Fundamental to the platforms' business model, algorithms are taught to manipulate users' brain chemistry in order to maximize their time online. What this often results in is an alteration of user worldview and behaviors, because, as we know, algorithms amplify content built on outrage, hate, and harmful material that generates more user engagement.

A clear example of the development of AI designed to alter individuals' emotional states through the delivery of information is Facebook's 2012 experiment on mood alteration through curation of news feeds.[18] This is connected to their research on AI inferences about personality type through Facebook 'Likes'.[19] The Cambridge Analytica scandal with its use of behavioural micro-targeting techniques to profile and target voters in a bid to influence voter behaviour is an indication of the way this type of AI can have very serious societal consequences as well as an impact on individual rights. And the leak of Facebook documents in Australia in 2017[20] which showed Facebook was selling insights into teenagers' emotional states in real time for targeted advertising is another indication of the way this kind of technology can impact on vulnerable groups, including children, by trying to access their inner states.

Equality of treatment

Much has been said about the manner in which AI is flawed because of the inherent bias built into the algorithms, linked to the lack of diversity of participation and opportunity in the industry that designs the algorithms. **Related to these are concerns about the lack of equal treatment in facilitating the inclusion of all users, and in monitoring for unequal impact on all users.** Through Avaaz's own investigation into hate speech on Facebook against communities of poor Muslims in the northeast state of Assam in India, we learned that AI is not an equal-opportunity capability -- indeed, it actively discriminates against some of the most vulnerable populations in the world.[21]

How are these rights engaged?

Through our investigation, we found that machine learning is not sophisticated enough, without proactive human-led content reviews, to extract hate speech from platforms, particularly in languages that are not very widely spoken. The danger of this , of course, was that this was the case despite three UN letters sounding the alarm bells about an emerging humanitarian crisis in Assam. Translation tools did not extend to these languages. But more fundamentally, the deployment of AI tools in the domain of hate or dangerous speech rests on a faulty premise: that all users have equal access to the flagging mechanism on Facebook's platform. Automated detection can only begin to function when there are an adequate number of posts flagged in the first instance from which classifiers can be built, or in simpler terms: humans need to flag content to train Facebook's AI tools to detect hate speech on its own. But, often, the minorities most directly targeted by hate speech on Facebook often lack online access or the understanding of how to navigate Facebook's flagging tools, nor is anyone else reporting the hate speech for them. As a result, the predictive capacity of AI tools is not equally robust.

[18] A.D.I. Kramer, J.E. Guillory, and J.T. Hancock Experimental evidence of massive-scale emotional contagion through social networks (2014) issue 24 of Proc Natl Acad Sci USA (111:8788–8790)

[19] W. Youyou, M. Kosinski, D.Stillwell ,Computer-based personality judgments are more accurate than those made by humans (2015) Vol. 112 No. 4, PNAS 1036-1040.

[20] https://www.theguardian.com/technology/2017/may/01/facebook-advertising-data-insecure-teens

[21] https://avaazpress.s3.amazonaws.com/FINAL-Facebook%20in%20Assam_Megaphone%20for%20hate%20-%20Compressed%20(1).pdf

International corporate accountability principles require platforms to conduct human rights due diligence on all products, such as identifying its impact on vulnerable groups like women, children, linguistic, ethnic and religious minorities and others, particularly when deploying AI tools to identify hate speech, and take steps to subsequently avoid or mitigate such harm. Ultimately, platforms need to be able to implement their policies equally for all populations, including vulnerable populations, so that hate speech can be accurately classified, identified, labelled, downgraded and removed quickly.

As the High-Level Expert Group on AI has stated "Bias and discrimination are inherent risks of any societal or economic activity. Human decision making is not immune to mistakes and biases. However, the same bias when present in AI could have a much larger effect, affecting and discriminating many people without the social control mechanisms that govern human behaviour."

Data Rights

The AI Framework must keep up with and anticipate the rapid industry developments in the terrain of content delivery. The conceptual framework must expand beyond data rights concepts of consent to use of data given to cover use of data created or inferred during AI automated decision making.

How are these rights engaged?

This data use creates repetitive patterns sending users down radicalization rabbit holes, draws users into filter bubbles and echo chambers that narrow their exposure, and promotes addictive behaviors, particularly in younger users who are more susceptible to the effects of disinformation. It thus becomes clear that **the harm of the unregulated algorithm is its potential to interfere with human autonomy: our personal data is being extracted to draw hidden inferences about us, which then allows our thoughts and emotions to be manipulated.**

We can see the tragic outcome of AI driven content curation without regulation in the story of UK teenager Molly Russell. Molly was just 14 when she took her own life.[22] After Molly died in 2017, her family looked into her Instagram account and found "bleak depressive material, graphic self-harm content and suicide encouraging memes. Her father believes this social media encouraged her desperate state, and described the process clearly: "Online, Molly found a world that grew in importance to her and its escalating dominance isolated her from the real world. The pushy algorithms of social media helped ensure Molly increasingly connected to her digital life while encouraging her to hide her problems from those of us around her, those who could help Molly find the professional care she needed."[23]

[22] https://www.bbc.co.uk/news/av/uk-46966009/instagram-helped-kill-my-daughter

[23] Ian Russell, Molly Russell's father in his forward to the report on technology use and the health of children and young people, from the Royal College of Psychiatrists in 2019 see

Recently, the Royal College of Psychiatrists has called on social media companies to share data with researchers to measure mental health impacts on young people of microtargeting, filter bubbles, and advertising.[24]

Product liability

We have noted that the Framework will consider issues of AI product liability under the current EU product safety and liability framework, however the AI Framework must specifically consider how victims can get redress from the kinds of AI decision making we have outlined above. Algorithms are products which can cause material and immaterial (e.g. loss of privacy, discrimination, violations of fundamental rights and human dignity) harm. More specifically, as explained above, the way algorithms are designed can amount to a violation of people's freedom of thought, which is an immaterial harm. It will be difficult to prove defect, damage, and a causal link between both.[25]

We welcome the EU Commission's openness to amending the product liability and safety framework to better accommodate redress for harm caused by AI. With regard to algorithms, we understand that some options being considered[26] are increasing transparency requirements to address the problem of opacity of systems based on algorithms, as well as requirements for robustness, accountability, human oversight and unbiased outcomes.

Avaaz fully supports such reforms and has developed policy proposals[27] which should be part of a comprehensive product liability and safety reform with respect to algorithms. We explore further how algorithms could be regulated within the AI Framework to ensure transparency, robustness, accountability, and human oversight requirements to make platforms "detox their algorithms" in Appendix 3 of this submission. A clear "detox the algorithm" obligation enshrined in a binding regulation can give victims of algorithmic harm a legal avenue to obtain redress.

https://www.rcpsych.ac.uk/docs/default-source/improving-care/better-mh-policy/college-reports/college-report-cr225.pdf

[24] ibid

[25] Note in the regard the complexities and delay in the UK court's consideration of redress for unauthorised data collection in Lloyds v Google LLC. The decision in the lower courts that "it was difficult to calculate exactly how many people had been affected and claims they had suffered damage were not supported by the group bringing the case" was eventually overturned on appeal but only after a lengthy and expensive class action.

[26] https://ec.europa.eu/info/sites/info/files/report-safety-liability-artificial-intelligence-feb2020_en_1.pdf

[27] https://secure.avaaz.org/campaign/en/disinfo_legislative_principles/#principles-list

Appendix 3

An audit framework for AI

The EU's AI framework must keep up with and anticipate the rapid industry developments in this terrain. The transparency that effective future regulatory oversight will need will depend on audit - both self audit, and audit if required by external regulators or trusted third party researchers.

The aim of an algorithmic audit for those deploying AI in their data processing should be to assess the impact on rights so that they can be protected appropriately, to identify mitigation strategies for unintentional and intentional harms, and specifically reduce the spread of harmful misinformation. This is equally relevant in the development and the deployment of AI in a range of sectors, including media and information platforms. We believe Avaaz's audit recommendations below have specific benefits in relation to AI data processing that results in any form of content moderation, curation, selection or recommendation.

The audit should cover:

- The purpose of the AI
- An assessment of the rights that could be impacted by the use of the AI
- The design of the algorithm to prevent rights breaches and mitigate risks to well-being
- Audit should supply evidence showing steps taken to facilitate the exercise of individual rights to reject biased/inaccurate sources of information - for example identifying and downgrading down categories of state sponsored media known to have published verifiable disinformation on a regular basis, known conspiracy domains, etc.
- Impact of the algorithms' operation on vulnerable populations such as racial and religious minorities, elderly, children, people with addictions, etc.
- The impact of the algorithm on the exercise of individual rights with sufficient clarity to the user to allow meaningful choice as to whether to engage with the service.

The Audit Process

When we think of regulatory oversight and the kind of questions that should be asked, we really need to think about it as a two step process - planning and audit of the design of the AI and its code in the context of its likely usage, and then periodic audit of its output - how it functions out in the real world in its ability to detect and deal with disinformation or other rights abuses resulting from its data processing. This in turn should lead to the identification and mitigation strategies for unintentional and intentional harms.

1) Design

Any Audit during the design and start up phase of an AI tech user should evidence the consideration given to the potential impact on the full range of human rights in the ECHR and the EU Charter. This should include the lawfulness of the purpose of the algorithm, as well as the risks inherent in the AI such as whether an algorithm's dataset is broad enough to be representative of all the conditions the system is likely to encounter for example, to mitigate the inherent bias against minority groups whose culture and language are not included in the data sets from which the algorithm learns.

Specifically Avaaz recommends that platforms using open algorithmic recommendation intended to serve a significant user base should provide evidence on whether the effect of a given algorithm on users rights and well being has been properly assessed and anticipated, with safeguards put in place to protect those rights, **by design**.

This audit should assess:

- **The impact of the algorithms' operation on its users well being** and whether sufficient controls within the algorithm have been designed to detect and mitigate risks during operation, for example data collection beyond the reasonable interpretation of consent that allows profiling and what behaviors such data collection may be promoting, for example addictive behavior or increasingly marginal, dangerous or polarised content as in the case of Molly Russell.
- **The algorithm's transparency:** Do users understand what data the algorithm is obtaining from them, how that data is being used, to whom that data is being provided?
- **The extent to which the algorithm facilitates the exercise of individual rights,** including the rights related to automated decision-making. Does it give sufficient explanation to users to enable them to understand choices made by algorithms that affect their human rights and freedoms? This is essential in the recommendation and search services of social media platforms.
- **Is the AI sophisticated enough to undertake the function it is deployed for?** For example if its function is to moderate content by automatically monitoring users data in terms of their user generated comments, has it been given sufficient language models to detect illegal speech, particularly hate speech against minorities? This not only requires the whole spectrum of relevant detailed language models but also data that allows learning of particular cultural nuances. It needs to pick up the pattern of usage that identifies an offensive racial slur in a given context. This is more than possible if sufficient attention is given to the algorithm's design. We are aware for example that Disney's algorithms on its children's services are capable of recognising when non insult words such as "chair" are used to insult users. By contrast, we reported on the failure of Facebook's algorithms or human moderators to pick this up during the eastern Indian state of Assam's drive to detect so-called illegal immigrants - for example the term

"Miya", which was originally an inoffensive term, is now a pejorative word to refer to Muslims in Assam. This usage of the term was not recognised as Hatespeech by the algorithm, and the class it was aimed at are generally too economically deprived to have access or knowledge to use Facebook's flagging tools.[28]

- **Is the algorithm's dataset broad enough** to be representative of all the conditions the system is likely to encounter? This both can mitigate potential assumptions and initial potential internal biases built into the algorithm and or that develop during usage to ensure the algorithm does not work to exacerbate existing biases in society.[29]
- **Whether the algorithm's dataset includes ongoing feedback mechanisms** to constantly improve it to **detoxify the algorithm** in order to support user's wellbeing and facilitate the exercise of their rights, allowing it to learn from the context in which it operates. This feedback should include the identification of disinformation - whether through the algorithm itself or through user reports - and lead to efficient correction of disinformation[30]. It should also be able to support effective anti-disinformation programmes including:
  - *Extracting from and/or devaluing all content identified as misinformation from a content curation or recommendation engines.*

AI should be able to learn the difference between the value of truthful content and misinformation categorized as such by reputable fact checkers. By designing algorithms that extract content identified as misinformation from their recommendation engines, the platforms can significantly reduce its spread while also respecting freedom of expression. Content creators would have freedom of speech, not freedom of reach.

- *Three strikes rule for misinformation*

Ie if a channel is detected to have spread misinformation, or violated the platforms guidelines in an effort to spread misinformation, deceive users or manipulate the algorithm more than three times, the **channels' content should not be further promoted through the platforms recommendation engine**. Channels should of course be given an opportunity to issue corrections to their viewers or challenge the decision if they disagree with it, but those channels unwilling to stop their malicious behavior should not be amplified.

A three strikes rule tackles misinformation while preserving freedom of expression -- content isn't deleted, but AI is designed with capacity to recognise it and no longer boost

[28] See MEGAPHONE FOR HATE: DISINFORMATION AND HATE SPEECH ON FACEBOOK DURING ASSAM'S CITIZENSHIP COUNT https://avaazpress.s3.amazonaws.com/FINAL-Facebook%20in%20Assam_Megaphone%20for%20hate%20-%20Compressed%20(1).pdf

[29] This would seem particularly relevant to the guidance you give on managing inherent bias in the AI at page 52 and automation bias at p99 of the Guidance.

[30] For more detail on Avaaz's "Detox the algorithm" policy proposal please see our previous submission to your consultation in response to the Alan Turing and ICO Guidance to AI users.

it. With time, this policy will help ensure that high quality content is more prominently promoted by the algorithms, while misinformation actors are marginalised.

- *Demonetize misinformation actors*

Under no circumstances should disseminators of misinformation be paid by for deliberately misleading content. Whilst this is less a question of AI design, than the appropriate action a platform should take when its AI identifies misinformation, we believe as a principle the issue of demonetising false and misleading information and data should be part of the AI's audit framework goals

1) Auditing and Transparency during operation

The Audit data should provide evidence to enable assessment of the impacts on the rights and freedoms and well being of individuals during the algorithm's operation and consider and record potential human rights impacts beyond data protection and privacy as a matter of course. Audit should evidence the ability of algorithms to mitigate risks detected, and give a clear account of any trade-offs as between rights - for example the trade off as between privacy and freedom of expression.

If the audit is going to be able to assess the public interest impacts of the AI as it operates, it must be conducted regularly during operation. We have laid out below the range of audit measures we believe are required for all responsible operators who manage open AI content recommendations services, and do so through the use of the data of their users. Over time such audits will expose an algorithm's unintended outcomes, we are sure that the social media giants did not intend tier AI to be used to misinform or troll their users, audits could have pointed to adaptations that could have corrected their course earlier.

An audit designed specifically to tackle an identified risk of disinformation should be able to assess the output of the algorithm against the following measures:

- The scale of disinformation directed to users though the AI's data processing of the user's data including:
  - the number and frequency of user reported breaches of platform standards, specifically reports on hate speech and disinformation;
  - The number and frequency of disinformation detected through the platform's moderation algorithms and/or reported by reputable fact-checkers;
  - The reach of disinformation on the platform - every platform can model the reach of a particular piece of content and this data should be provided to the auditors:
  - The numbers of fake accounts detected and removed;
  - The extent of unlabelled bots on the platform
  - Other patterns of inauthentic behaviour.

- The efficacy of the platform’s algorithms to detect and mitigate breaches of the platforms standards. This would include for example:
    - The **speed** with which all reports are **assessed** by the platform’s moderation algorithms and or human moderators;
    - The **speed** at which a **correction** is placed on misinformation content, and a comparison between the reach of the misinformation and the amount of views on the correction.
    - The **nature of any other action** taken in response to all reports, and the speed with which it was taken;
    - The **reach** of any correction the platform provided alongside the data on the reach of a particular piece of content
    - The **degree** to which the control measures in the algorithm **downgraded and suppressed promotion** of a given piece of disinformation;
    - Any further action taken in respect of the piece of disinformation content or its source - such as the demonetisation of the content or the channel on which it was spraread
    - The **removal of account**s spreading illegal material such as hate speech;
    - The **ability of the algorithm to detect repeat attempts** by such account holders to game the system by creating new accounts;
    - In the case of a user reported breach, the **communication of action taken to the user** who reported it; and
    - Where a breach could affect the rights and/or well-being of a wide set of users **on issues of public interest -** for example content that stirs up hatred against immigrants, false claims about a crucial story in an election, or bogus claims on a cure for Covid-19, then **communication of the breach and the action taken** against the account holder’s who created it, should be provided to **all affected users within the platform**, not just to the person who reported it. This is the only way the public can gain insight into the scale and organisation of disinformation on the platform. Our full policy position on Correct the Record along with a study showing the effectiveness of corrections disseminated to all who were exposed to disinformation can be found here: Avaaz White Paper: Correcting the Record https://secure.avaaz.org/campaign/en/correct_the_record_study/