Avaaz Response to the Special Rapporteur's request for input into the annual thematic report to be presented to the Human Rights Council at its 47th session in June 2021.

Ms. Irene Khan
Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression
OHCHR-UNOG
8-14 Avenue de la Paix
1211 Geneve 10, Switzerland
By email freedex@ohchr.org

15 February 2020

Dear Madam

Avaaz welcomes this renewed focus on Disinformation from the UN Special Rapporteur and her team. Avaaz has over 66 million members worldwide who are passionately invested in combatting the harms to our human rights posed by disinformation online. Avaaz has been investigating and campaigning on the threat posed by disinformation, organized and distributed at scale by AI on social media platforms for over two years. We have published authoritative reports based on hard data on its impact on the whole range of human rights[1]. We have been in detailed legislative conversations advocating pragmatic and exper-based approaches to solve the menace of disinformation.

Avaaz's own research provides ample evidence of the impacts on human rights of disinformation across the globe.

A snapshot of our reports in 2020 shows disinformation with direct impact on the right to life: for example:

- Between June and September 2020, a study by Avaaz found that at least 50 pages and 27 public groups on Facebook shared content glorifying violence, praising Kyle Rittenhouse, or spreading misinformation. Those pages and groups amassed 19.2 million followers in total, and garnered over 114 million interactions.

- In the run up to the US Elections Avaaz found disinformation narratives promoted by a network of nine Facebook pages working in a coordinated manner to amplify and potentially monetise the virality of outrage. These narratives attained at least 26 million estimated views combined; and included that:
    - Black Lives Matter activists are threatening to assassinate white families;
    - The anti-racism protests are using paid protesters who have been hired through the website ProtestJobs.com;

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

- "Antifa" groups or government officials placed pallets of bricks at protest sites in US cities to stoke violence;
- George Floyd is alive.

- In August 2020, we reported on content from the top 10 websites in the EU spreading health misinformation had 470 million estimated views on Facebook -- almost four times as many as equivalent content from the websites of 10 leading health institutions, such as the WHO and the CDC.

We can't hope to cover the depth of our research and engagement in this arena in a letter and believe it will be most productive to arrange a meeting with you and your team during your analysis period so we can supplement your understanding, and answer any questions you may have as you continue your investigation.

To give you an idea of the range of expertise we have we have attached the following appendices:

1) An analysis of the rights, freedoms and values impacted by disinformation disseminated through social media;
2) Current Disinformation threats across the Globe - Avaaz reports summaries for the past two years;
3) Avaaz proposals on countering disinformation by **correcting the record,** which is a means of providing transparency and more information to users, and initial studies indicate this solution could decrease belief in disinformation by an average of 50%**;**
4) Avaaz proposals to make platform's accountable for their algorithms, and stop the spread of disinformation **through detoxing the algorithm,** which ensures that harmful disinformation is not amplified by algorithms; and
5) An audit framework to address disinformation on social media platforms.

Our full reports can be found at https://secure.avaaz.org/campaign/en/disinfo_hub/ but we do urge a meeting so we can talk through our solutions and further unpublished data on, for example, the role of algorithms in amplifying disinformation during the US2020 elections. We have yet to release our final analysis, but our preliminary findings show how simple changes Facebook made to its algorithm the weeks leading to the election, and the weeks after, significantly impacted the proportion of engagement on authoritative content versus disinformation content.

We would also like to speak to you about our recommendations on a code of best practice, which we are developing to help the design of a new European Digital Framework to combat disinformation, and which models the practical paths we see to resolve this issue globally. We look forward to meeting with you and your team very much,

Yours sincerely

Sarah Andrew and Meetali Jain
Legal Directors of Avaaz Foundation, Anti-Disinformation Project and Avaaz Foundation respectively
Contacts:
Sarah.Andrew@Avaaz.org and **Meetali@Avaaz.org**

Appendix 1

The rights, freedoms and values impacted by disinformation disseminated through social media

The Right to Life

Allowing hate speech to proliferate that incites harm to one's safety or security of person, or misinformation that exposes individuals to significantly elevated risks to their health arguably violate a corporate entity's duty to respect the human rights to life and health. Our analysis of hate speech in Assam revealed dangerous levels of inciteful lies and hate, and in several instances, spread by state agents, against Muslims in Assam. We also found a massive amount of misleading or false information about covid-19 and vaccines that could be, and in some instances, has been relied upon by individuals to harm their health.

Freedom of Speech / Freedom of Expression

It is clear that laws that contain blanket bans on misinformation or untruthful speech infringe on the freedom of expression. But in certain instances spreading misinformation can deny individuals' right to seek and receive accurate information through what is termed "censorship through noise". Disinformation has also been used to target or troll journalists and human rights defenders critical of governments or political movements. These online harassment campaigns can produce a chilling effect on the freedom of expression, association and assembly of the targeted individuals, who may refrain from publicly expressing their views and engaging in their normal activities for fears of further verbal and physical attacks.

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 deceptive amplification of propaganda and microtargeting 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 untransparently 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.[2] This is connected to their research on AI inferences about personality type through Facebook 'Likes'.[3] 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[4] 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. This has been termed algorithmic bias or algorithmic determinism,** 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.[5]

How are these rights engaged?

---

[2] 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)

[3] 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.

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

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

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.

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 current data rights concepts of consent to user rights to understand, control and actively choose the degree to which they are micro targeted or surveilled through use of their own data as well as 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.[6] 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.”[7]

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.[8]

[6] https://www.bbc.co.uk/news/av/uk-46966009/instagram-helped-kill-my-daughter
[7] 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 https://www.rcpsych.ac.uk/docs/default-source/improving-care/better-mh-policy/college-reports/college-report-cr225.pdf
[8] ibid

Appendix 2

Current Disinformation threats and Avaaz reports summaries between 2019 and 2020.

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

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.**

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.**

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.

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.

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.

Anti-Racism Protests: Divisive disinformation narratives go viral on Facebook 12/9/2020
https://secure.avaaz.org/campaign/en/anti_protest_disinformation/
As millions of people throughout the world hit the streets to protest police brutality and racism, Avaaz's investigative team found disinformation seeking to polarize the debate, or to demonise and undermine the protests spreading and destabilising the debate. The report tracking the disinformation live across the period estimates that a small sample of posts connected to a dozen viral disinformation narratives about the anti-racism protests had been viewed millions of times on Facebook during that time.

BRIEF: How Facebook's AI is failing American voters ahead of Election Day October 2020
https://secure.avaaz.org/campaign/en/facebook_fact_check_failure/

Our report details the loopholes in platform's attempts to stem aariations of misinformation already marked false by Facebook. These included for example content alleging Presidential Candidate Joe Biden is a paedophile, that President Trump's family stole money from a kids charity, or that multiple stamps are needed for mail-in ballots. Clones of this information, which facebook knows to be incorrect through its fact checking partner network, - are slipping through Facebook's detection system and being viewed millions of times ahead of the elections, according to Avaaz.

While Facebook said it would apply 'strong warning labels' to fact-checked content and reduce its distribution, Avaaz's new investigation found that nearly half of the fact-checked misinformation

content (42%) we analysed is managing to circumvent Facebook's own policies and remain on the platform without a label, earning millions of interactions.

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.

FACEBOOK'S ALGORITHM: A MAJOR THREAT TO PUBLIC HEALTH August 19, 2020 see
https://secure.avaaz.org/campaign/en/facebook_threat_health/
Avaaz uncovers health misinformation spreading networks with an estimated 3.8 billion views in 2020 - and shows how to quarantine this infodemic. Avaaz investigated the leading misinformation claims including for example that in 2017, the WHO admitted that the global explosion in polio is predominantly caused by vaccines or misinformation misreporting on a case regarding HPV vaccinations in the Indian Supreme Court making allegations about the Gates Foundation. The content we reported on from the top 10 websites in the EU spreading health misinformation had **470 million estimated views on Facebook -- almost four times as many as equivalent content from the websites of 10 leading health institutions, such as the WHO and the CDC.**

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 3

COUNTERACTING DISINFORMATION: CORRECTING THE RECORD

Independent studies have shown that disinformation has become such an unprecedented threat to our democracies, because on social media false information spreads up to six times faster than the truth. So even if fact-checked and found untrue -- the millions of people who have seen the false content in the first place will likely never find out that they have been misled.

Governments protect consumers and participants in financial and energy markets from false and misleading information, including by making it possible to issue corrections when misinformation could influence people's decisions. We should offer our democracies the same protections we offer our Markets.

The solution is simple: platforms themselves must inform users and push effective corrections to each and every person who saw the false information in the first place. Newspapers publish corrections right on their own pages, television stations on their own airwaves; platforms should do the same on their own channels. No one else can do it.

Corrections work: Multiple peer-reviewed studies have demonstrated that effective corrections can reduce and even eliminate the effects of disinformation. Studies attempting to replicate the often discussed 'backfire effect' -- where corrections entrenched false beliefs -- have instead found the opposite to be true. Meanwhile, researchers are converging best practices for effective corrections. In our view, correcting the record would be a five-step process:

1. Define: The obligation to correct the record would be triggered where:
   ○ Independent fact checkers verify that content is false or misleading;
   ○ A significant number of people -- e.g. 10,000 -- viewed the content.
2. Detect: platforms must:
   ○ Deploy an accessible and prominent mechanism for users to report disinformation;
   ○ Provide independent fact checkers with access to all content that has reached e.g. 10,000 or more people.
3. Verify: Platforms must work with independent, third-party verified fact-checkers to determine whether reported content is disinformation, as defined by the EU, within 24 hours.
4. Alert: Each user exposed to verified disinformation should be notified using the platform's most visible notification standard.
5. Correct: Each user exposed to disinformation should receive a correction that is of at least equal prominence to the original content, and that follows best practices which could include:
   ○ Offering reasoned alternative explanation, keeping the users' worldview in mind, emphasizing factual information while avoiding, whenever possible, repeating the original misinformation;

Appendix 4

STOPPING THE SPREAD OF DISINFORMATION BY DETOXING THE ALGORITHM

Social media companies' 'curation algorithms' decide what we see, and in what order, when we log on. They're designed to keep us glued to the screen and always wanting to come back for more. They succeed in part by pushing emotionally charged, outrageous and polarizing content to the top of our feeds. That's one of the big ways hatred, disinformation, and calls to political violence go viral.

Fortunately, this can be fixed. Having designed and developed them, platforms can Detox their Algorithms by making sure known disinformation is transparently downgraded, not amplified, in our feeds, by demonetizing disinformation, and being transparent with their users by using alerts.

Facebook's own research shows that slowing the spread of disinformation can reduce views by up to 80%.But this solution is not being rolled out at scale at Facebook or other major social media platforms.

Research shows that curation algorithms can lead 'regular people' to extremism. An internal report at Facebook in 2016 revealed that 64% of users who joined an extremist group on its platform only did so because the algorithm recommended the groups. One study demonstrates that YouTube's recommendations prioritizes extreme right-wing material after interaction with similar content.

Three Steps to Stop the Spread and Detox the Algorithms:

1. Detect and downgrade with full transparency known pieces of misinformation and all content from systematic spreaders. All platforms should stop accelerating any content that's been debunked by independent fact-checkers, as well as all content from pages, groups, or channels that systematically spread misinformation. Users whose content, pages, groups or channels are

2. Demonetize systematic spreaders. When an actor has been found to be systematically posting fact-checked content, the platforms must ban these actors from advertising and from monetizing their content.

3. Inform users and keep them safe. Users should be informed through clear labels when they're viewing or interacting with content from actors who were found to be repeatedly and systematically spreading misinformation, and be provided with links to additional information.

Detox the Algorithm protects free speech by requiring that all content remains available and guarantees users due process -- the right to be notified and to appeal the platforms' decisions. It also protects freedom of thought by slowing the spread of harmful lies that change how our brains are wired.

Appendix 5

An audit framework address disinformation on social media platforms

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.

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.[9]

[9] 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

- **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.[10]
- **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[11]. It should also be able to support the effective anti-disinformation programmes laid out in Appendix 4 of this document.

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 which accelerates the content on the social media platforms as they operate, it must be conducted regularly during operation. We have laid out below a range of structural and service level 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:

10

11

- 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/