Avaaz Position Paper on the Digital Services Act, Disinformation and Freedom of Speech

The Digital Services Act (DSA) is Europe's historic opportunity to address key issues endangering democracies at a global and unprecedented scale. It attempts to strike a careful balance by restricting content removal - which could impinge on the freedom of speech - only to illegal content, whilst ensuring that the full range of impacts on our fundamental rights are dealt with through a procedure of risk assessment and mitigation. We see DSA's role as laying down a comprehensive approach for this digital century that provides an overarching framework to address:

1) Clear liability for hosting or disseminating illegal content, with the removal of such content on a "notice and take action" basis

2) Auditable, meaningful transparency provided to users, researchers and regulators

3) Procedural rules to assess and mitigate the massive societal damage that harmful but legal content such as disinformation can achieve once accelerated to millions through the platform's algorithms

4) Giving power back to consumers with meaningful choice over the algorithms that determine the way in which we see and experience our digital lives

5) Providing a realistic co-regulatory backstop to the commitments promised by platforms under the Code of Practice on Disinformation (CoP)

All Avaaz policy proposals are based on the experience of over three years of research on disinformation and other harmful content. Avaaz has been one of the leading organisations in the EU and around the world, reporting on disinformation and its scale, commissioning polls on its harmful effects, evaluating platforms' efforts to manage it, and identifying their failures, with platforms analysed including Facebook, Youtube and

WhatsApp. More recently our research showing how harmful but legal content on social media has created the ecosystem which allowed the Capitol Hill insurrection to happen was used by the US Congress to question tech CEOs.

Our position is that social media platforms are not mere content hosts which just provide a neutral "content-blind" service. Nor are they content publishers with editorial responsibility for the content users upload. They are **content accelerators** responsible for the amplification of content to millions of people. Their business model is based on algorithms using vast amounts of data collected from its users to decide which content to serve us, with the main purpose of keeping us on their platform for as long as possible.

As ERGA said in its statement on the Digital Services Act and Digital Market Act proposals[1], "a fully-fledged and consistent approach to online content platforms' moderation of harmful content is (...) necessary." Here are our proposals:

Auditable, meaningful Transparency for All: users, researchers and regulators

The full transformative power of transparency is unlocked when meaningful transparency is provided to **users, researchers and regulators.**

In that sense, we welcome the current transparency measures in the text, especially the perspective of transparency from the platforms as a basic right to the users (Articles 12, 14, 15, 17). It will open and shed light on the decision-making process of the platforms on content moderation.

Transparency to users - Amending Article 15

Some improvements are necessary. Article 15 is currently drafted to ensure platforms provide statements of reasons to the producers of content where that content has been removed from view. This is an important step, but it must also include other possible moderation actions a platform has taken, either as a result of notice and take action mechanisms under Article 14, or as a result of the risk assessment and mitigation measures under Articles 26 and 27.

Recent investigations have shown how platforms' decisions to act or not on specific content or actors are highly non-transparent, inconsistent, and have the potential to influence the political debate. As many of these decisions happen away from public scrutiny, what the additional transparency regulation would bring is not a threat to freedom of speech – it actually is needed to preserve it.

[1] https://erga-online.eu/wp-content/uploads/2021/03/ERGA-DSA-DMA-Statement_29032021.pdf, page 4

The right to know about the actions a platform has taken extends beyond the content creator to those exposed to the illegal and/or (in the case of risk assessment and mitigation measures) legal but harmful content which is not removed but moderated in some other way.

Users have the right to know when they have been a target of a fake account, an influence operation, or platform manipulation and who targeted them. That way, users are going to be empowered to understand which kind of content and actors they are served, fostering citizens' media literacy, increasing their understanding of the phenomenon of disinformation and increasing their resilience to it.

One example among many can be drawn from an Avaaz investigation in May 2020: Natural News, at that time the biggest known health- and covid-misinformation spreading website in the world, with more interactions in a year than the WHO and CDC websites combined, was deplatformed by Facebook because of a violation of Facebook's terms of service - Natural News had used troll farms in North Macedonia.

It was a hugely popular website, and from one day to the next, all their pages and groups on Facebook were removed. Users trying to post links from the website were notified with a pop-up saying "You can't share this link. Your post couldn't be shared, because this link goes against our Community Standards".

Because the platform did not share any information about *why* Natural News was deplatformed, people started claiming this was social media censorship, were deeply unhappy with the platform's actions, and never knew that they were targeted by a malicious actor.

We therefore suggest that Article 15 is amended as follows[2]:

Current Draft	Suggested amendment
Article 15 1. Where a provider of hosting services decides to remove or disable access to specific items of information provided by the recipients of the service, irrespective of the means used for detecting, identifying or removing or disabling access to that	Article 15 1. Where a provider of hosting services decides to remove or disable access to, **or otherwise moderate either the form or distribution of** specific items of information provided by the recipients of the service, irrespective of the means used for detecting, identifying or removing or disabling access to, **or otherwise moderating**

[2] NB throughout this document changes and additional text to Articles are suggested in bold text

information and of the reason for its decision, it shall inform the recipient, at the latest at the time of the removal or disabling of access, of the decision and provide a clear and specific statement of reasons for that decision.	that information and of the reason for its decision, it shall inform the recipient, at the latest at the time of the removal or disabling of access, **or other moderation measure,** of the decision and provide a clear and specific statement of reasons for that decision.
2. The statement of reasons referred to in paragraph 1 shall at least contain the following information:	2. The statement of reasons referred to in paragraph 1 shall at least contain the following information: [...] (f) **Where the decision is based on an assessment of a risk under the risk assessment procedure of Article 26 of this Act, a reference to the risk identified and explanations as to why any mitigation measure applied under Article 27 of this act was considered to be required to mitigate that risk.**

Algorithmic transparency and algorithmic choice - Amending Article 29

We were pleased to see the ideas within paragraph 62 of the recitals to the DSA, which proposes that users of very large online platforms should be informed, and be able to influence the information presented to them. It suggests that platforms should clearly present the main parameters for their recommender systems to users in an easily comprehensible manner to ensure that the recipients understand how information is prioritised for them.

We see this as a key mitigation measure. It's through understanding the nature of the service they are using, and its vulnerabilities to disinformation, that users are empowered to make real choices on how they want the service's algorithms to serve them content, to try to ensure disinformation reaches their social media feeds less and less.

Similarly users must be afforded the right to be fully informed about the kinds of data collection used to target them with the content they see. A revolution on the culture of data collection is needed, so that platforms stop seeking a waiver of rights through a tick box to enter the service but work to build their user's understanding on what categories of data are stored, what uses the data is put to, how tracking of behaviours and data aggregated from other services inform the algorithm's selections.

Accordingly, **Article** 29 (1) should provide for - not merely recommend - active user choice over which algorithm serves them, and compel platforms to provide users with enough information to make a meaningful choice. Users should also have the right to reject algorithms based on user profiling, and the option to turn off algorithmic selection, or to choose which elements of it they wish to enable, just as they have rights to tailor the data collection they allow under GDPR, similar to the cookies opt-out.

Current Draft	Suggested amendment
Article 29 1. .Very large online platforms that use recommender systems shall set out in their terms and conditions, in a clear, accessible and easily comprehensible manner, the main parameters used in their recommender systems, as well as any options for the recipients of the service to modify or influence those main parameters that they may have made available, including at least one option which is not based on profiling, within the meaning of Article 4 (4) of Regulation (EU) 2016/679.	Article 29 1. Very large online platforms that use recommender systems shall set out in their terms and conditions, in a clear, accessible and easily comprehensible manner, the main parameters used in their recommender systems, and **they shall provide** options for the recipients of the service to modify or influence those main parameters that they may have made available, including at least one option which is not based on profiling, within the meaning of Article 4 (4) of Regulation (EU) 2016/679."

2. Where several options are available pursuant to paragraph 1, very large online platforms shall provide an easily accessible functionality on their online interface allowing the recipient of the service to select and to modify at any time their preferred option for each of the recommender systems that determines the relative order of information presented to them:	2. Where several options are available pursuant to paragraph 1, very large online platforms shall provide an easily accessible functionality on their online interface allowing the recipient of the service to select and to modify at any time their preferred option for each of the recommender systems that determines the relative order of information presented to them, **including any technically possible option to turn off algorithmic selection within the the recommender system entirely**.

User choice only really works if the platforms also provide users with the transparency measures we have described above, to give them a thorough understanding of the nature of the service they are using, and its vulnerabilities to disinformation.

Finally, there must be a clear linkage between the transparency measures and the European Data Strategy - which asks that companies limit their data collection to what's absolutely necessary for their business function.

Transparency to regulators - Amending Articles 13 and 33

On the side of reporting obligations from platforms, the Annual Reporting should contain all the content moderation measures platforms engage in, including the ones taken as a result of the application and enforcement of their terms and conditions. The granularity foreseen in the text now doesn't address the measures that would hold back disinformation, leaving a big chunk of this work up to the Code of Practice to tackle.

Article 13 should therefore be amended to include, in a non-exhaustive list, the following:

Current Draft	Suggested amendment
Article 13 1. Providers of intermediary services shall publish, at least once a year, clear, easily comprehensible and detailed reports on any content moderation they	Article 13 1. Providers of intermediary services shall publish, at least once a year, clear, easily comprehensible and detailed reports on any content moderation they

engaged in during the relevant period. Those reports shall include, in particular, information on the following, as applicable: [...]	engaged in during the relevant period. Those reports shall include, in particular, information on the following, as applicable: **(e) the total reach and views of fact-checked disinformation, confirmed by independent fact-checkers, or which breaches some other aspect of their own misinformation or disinformation policy;** **(f) the total number and frequency of user-reported breaches of platform standards on disinformation;** **(g) the volume and frequency of disinformation detected through the platform's moderation algorithms;** **(h) the total volume of disinformation detected through AI proportionate and actioned by human review proportionate to the total volume of disinformation on the platform.**

We see as positive the fact that platforms must publish the results of their risk assessments and the mitigation measures they've adopted (Article 33), but this is weakened by paragraph 3, which establishes that if the platforms decide that information in their transparency reports is too private or sensitive it may remove that information from its transparency reports. This drafting leaves too much room for platforms to wriggle out of offering transparency as required and we have **suggested wording that limits such reasons to ones involving** "significant vulnerabilities for the security of its servicor which may undermine public security or may harm recipients"**.**

Current Draft	Suggested amendment
Article 33 [...] 3. Where a very large online platform considers that	Article 33 [...] 3. Where a very large online platform considers that

the publication of information pursuant to paragraph 2 may result in the disclosure of confidential information of that platform or of the recipients of the service, may cause significant vulnerabilities for the security of its service, may undermine public security or may harm recipients, the platform may remove such information from the reports. In that case, that platform shall transmit the complete reports to the Digital Services Coordinator of establishment and the Commission, accompanied by a statement of the reasons for removing the information from the public reports.	the publication of information pursuant to paragraph 2 may result in the disclosure of confidential information of that platform or of the recipients of the service **and that this** may cause significant vulnerabilities for the security of its service, may undermine public security or may harm recipients, the platform may remove such information from the reports. In that case, that platform shall transmit the complete reports to the Digital Services Coordinator of establishment and the Commission, accompanied by a statement of the reasons for removing the information from the public reports. **The Digital Services Coordinator should be required to publish its decision as to whether it upholds the platform's decision giving the reasons for this. This decision must be subject to review by the relevant judiciary body. For the absence of doubt, mere commercial confidentiality shall not be a reason for a very large online platform to fail to disclose the relevant information.**

Transparency to researchers - Amending Article 31

As a final point which has been echoed in conversations with other organisations and researchers, we strongly believe that the pool of researchers with access to the information provided by the platforms is too restricted. To best understand how the platforms work, how they affect individuals and society, and how they can be made to pose fewer systemic risks, researchers must be provided with access to data – data that has largely been kept from them (despite promises). There has been widespread recognition of the role that investigative journalists and NGOs with research and reporting functions like Avaaz and others have contributed to the evidence base upon which the DSA now builds. Yet Article 31 only requires platforms to provide researchers access on the Commission's request; only if the researchers are 'vetted'; and only for the purpose of helping the Commission identify systemic risks. We need to extend the research rights to investigative journalists and civil society, as the platforms already have a right to refuse intelligence on grounds specified in Article 33. We also

suggest that as a matter of practice, although not to be specified in the DSA, that when those grounds are relied on to refuse access, a description of the ground, and a route of appeal should be given.

Current Draft	Suggested amendment
Article 31 [...] 2. Upon a reasoned request from the Digital Services Coordinator of establishment or the Commission, very large online platforms shall, within a reasonable period, as specified in the request, provide access to data to vetted researchers who meet the requirements in paragraphs 4 of this Article, for the sole purpose of conducting research that contributes to the identification and understanding of systemic risks as set out in Article 26(1). 4. In order to be vetted, researchers shall be affiliated with academic institutions, be independent from commercial interests, have proven records of expertise in the fields related to the risks investigated or related research methodologies, and shall commit and be in a capacity to preserve the specific data security and confidentiality requirements corresponding to each request.	Article 31 [...] 2. Upon a reasoned request from the Digital Services Coordinator of establishment or the Commission **or any body with proven records of expertise in the fields related to the risks investigated or related research methodologies**, very large online platforms shall, within a reasonable period, as specified in the request, provide access to data to vetted researchers who meet the requirements in paragraphs 4 of this Article**.** 4. In order to be vetted, **platforms may only require that** researchers shall have proven records of expertise in the fields related to the risks investigated or related research methodologies, and shall commit and be in a capacity to preserve the specific data security and confidentiality requirements corresponding to each request.

Risk assessment and mitigation

Assessing the risks: Amending Article 26(1), (b) and (c)

Articles 26 and 27 have the potential to bring meaningful regulation to risks posed by platforms that cannot be easily categorised in terms of the illegality of their content. They meet the need described by Commissioner Breton *"to set the rules of the game and organize the digital space with clear rights, obligations and*

safeguards."[3] The DSA rightly confines its ambitions to risks posed by the business models of the very large online platforms to our fundamental human rights. It is important, though, that those rights are not narrowly construed as being simply ones relating to freedom of expression.

The rights must be all rights protected by the Charter of Fundamental Rights, including the right to life, freedom of thought and opinion as well as the rights currently named in Article 26 of expression and information, the right to private life, the prohibition on discrimination and the rights of the child.[4] These rights may well require mitigation from harm which goes beyond the notice and take down regime for illegal content, to mitigate the harm of legal but harmful content where it impinges on our fundamental rights as EU citizens. As Commissioner Jourová said when she launched the Comission's work to combat disinformation, *"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".*[5]

This means that the risk assessment procedures under Article 26, (1), (b) must be deep and wide enough to cover all reasonably foreseeable human rights risks that the content acceleration or distribution systems of the very large online platforms could pose.

Current Draft	Suggested amendment
Article 26 1. Very large online platforms shall identify, analyse and assess, from the date of application referred to in the second subparagraph of Article 25(4), at least once a year thereafter, any significant systemic risks stemming from the functioning and use made of	Article 26 1. Very large online platforms shall identify, analyse and assess, from the date of application referred to in the second subparagraph of Article 25(4), at least once a year thereafter, any significant systemic risks stemming from the functioning and use made of

[3] https://www.politico.eu/article/thierry-breton-social-media-capitol-hill-riot/

[4] Please see Appendix 1 for a detailed analysis of the range of human rights impacted by algorithmic acceleration of disinformation content.

[5] https://www.politico.eu/article/european-commission-vp-backs-twitter-in-trump-battle/

their services in the Union. This risk assessment shall be specific to their services and shall include the following systemic risks: [...] (b) any negative effects for the exercise of the fundamental rights to respect for private and family life, freedom of expression and information, the prohibition of discrimination and the rights of the child, as enshrined in Articles 7, 11, 21 and 24 of the Charter respectively;	their services in the Union. This risk assessment shall be specific to their services and shall include the following systemic risks: [...] **(b) any negative effects for the exercise of the fundamental rights enshrined in the European Charter of Fundamental Rights;**

Additionally, the wording of Article 26, (1), (c) contains a worrying limitation to the assessment of the risks of a service. It states that the risk assessment is in relation to any **intentional manipulation of a service.** In fact, our investigations strongly show how the main source of service's vulnerability to the spread of harmful content, including misinformation, is the design of their content recommendation and monetisation systems and their own moderation systems.

A Wall Street Journal article[6] reported that in 2018 an internal meeting at Facebook outlined its understanding of how its 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". "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."

And so the speed and scale at which content "goes viral" is enhanced by the digital service's own design, 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. Therefore, there is a risk that users get trapped in an online bubble of disinformation, hate speech or other harmful content.

[6] https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions

The aim of disinformation, as Commissioner Jourová warned, "*to blur the lines, to polarise and make us indifferent*"[7] is massively amplified by the information processing capacity in the dissemination, recommendation and selection of media content. To fail to address the potential for risks which arise from the inherent design of the opaque data-driven algorithms would be a huge disservice to the potential reforming power of the DSA.

So, whilst it's arguable that the wording of Article 26 is not intended to exclude inherent design risks from the reach of the DSA, as clauses a - c are not exhaustive, it is equally not clear on its face and we therefore suggest the following amendment to Article 26 (1)(c)[8]:

Current Draft	Suggested amendment
Article 26 1. Very large online platforms shall identify, analyse and assess, from the date of application referred to in the second subparagraph of Article 25(4), at least once a year thereafter, any significant systemic risks stemming from the functioning and use made of their services in the Union. This risk assessment shall be specific to their services and shall include the following systemic risks: [...] c) intentional manipulation of their service, including by means of inauthentic use or automated exploitation of the service, with an actual or foreseeable negative effect on the protection of public health, minors, civic discourse, or actual or foreseeable effects related to electoral processes and public security.	Article 26 1. Very large online platforms shall identify, analyse and assess, from the date of application referred to in the second subparagraph of Article 25(4), at least once a year thereafter, any significant systemic risks stemming from the functioning and use made of their services in the Union. This risk assessment shall be specific to their services and shall include the following systemic risks: [...] c) **actual or foreseeable systemic negative effects on the protection of public health, minors, civic discourse, or actual or foreseeable effects related to electoral processes and public security, including but not restricted to the risk of the intentional manipulation of their service by means of inauthentic use or automated exploitation of the service.**

[7] From the Opening speech of Vice-President Věra Jourová at the conference "Disinfo Horizon: Responding to Future Threats" https://ec.europa.eu/commission/presscorner/detail/en/SPEECH_20_160

[8] Under the same arguments, recital 57 needs to be expanded to incorporate risks, irrespective of intent, where they are caused by the platform's own, even unintentional, amplification of harmful content.

Involvement of civil society - Adding Article 26, (3)

It is also essential that the terms of these risk assessments are not left to be defined by the platforms, nor audited by private contractors commissioned by the platforms. They must be published and subject to scrutiny by civil society and relevant regulators. Recital 59 introduces the concept that the platforms should not be the final arbiters of the form of assessment and mitigation measures stating that: "Very large online platforms should, where appropriate, conduct their risk assessments and design their risk mitigation measures with the involvement of representatives of the recipients of the service, representatives of groups potentially impacted by their services, independent experts and civil society organisations."

We believe this should not be a question of choice, but mandatory, and would suggest that Article 26 is amended with a final additional sub clause:

Current Draft	Suggested amendment
Article 26	Article 26 **3. The risk assessments must be designed with the involvement of representatives of the recipients of the service, representatives of groups potentially impacted by their services, independent experts and civil society organisations.**

Mitigating the risks - Amending Article 27, para 1, (a) (c) and (d)

We know that the platforms are already undertaking some forms of activity to mitigate the risks created by the algorithms, for example by downgrading actors and disinformation. But they are doing so without transparency, threatening free speech by effectively becoming the private arbiters of what can and cannot be said. Avaaz is wholly opposed to prior moderation, and to removal of any content other than illegal content, but whilst assessment and mitigation processes will bring much needed transparency, transparency alone will not do the work the DSA needs to achieve.

We applaud the acknowledgement on Article 27 that concrete steps in the design of the service may be needed, but are concerned that it may allow too much leeway for platforms to interpret its provisions narrowly, particularly if a restrictive view of the risks assessed under Article 26 is taken. If Article 27 fails to achieve appropriate and proportional mitigation actions to address the risks identified under Article 26 above, it drills a hole through the architecture of the DSA. We believe it should be amended to explicitly include the requirements for platforms to mitigate risks identified under clause 26 by progressively redesigning their content recommendation algorithms, to place less emphasis on the retention of attention and reduce the reach of legal but harmful content and systematic inauthentic actors. Platforms must prioritise improving the variables which underpin their content recommendation algorithms, and the data sets that inform them.

We also believe that transparency to users needs to be included as a fundamental risk mitigation measure. The need for platforms to inform users about harmful content or platform manipulations they've been exposed to is not just a huge incentive for them to decrease the impact of such content, but can also undo a lot of the damage. If we consider in particular disinformation, over the last 10 years, the science of debunking has overwhelmingly confirmed that well-designed labelling of disinformation which points the users to verified information is an effective tool for fighting disinformation.

In October 2020, dozens of the world's top scientists in this sector came together and, inspired by the IPCC Climate report, published the Debunking Handbook, which consolidated the results of dozens of scientific studies on the issue concluding that "recent evidence provides no reason to avoid debunking for a fear of a backfire effect". Because of this proven efficacy and wide consensus, it is imperative that the DSA includes specific reference to a need to inform users who have uploaded, viewed, shared or otherwise engaged with harmful content.

As with Article 26, we believe that the involvement of civil society should be explicitly mandated in the risk assessment articles, namely with an additional sub clause to Article 27 (1)

Accordingly, we believe that Article 27 (1) should be amended as follows:

Current Draft	Suggested amendment
Article 27(1) a) adapting content moderation or recommender systems, their decision-making processes, the	Article 27(1) a) adapting content moderation or recommender systems, their decision-making processes, the

features or functioning of their services, or their terms and conditions; b) targeted measures aimed at limiting the display of advertisements in association with the service they provide; c) reinforcing the internal processes or supervision of any of their activities in particular as regards detection of systemic risk; d) initiating or adjusting cooperation with trusted flaggers in accordance with Article 19; e) initiating or adjusting cooperation with other online platforms through the codes of conduct and the crisis protocols referred to in Article 35 and 37 respectively.	features or functioning of their services, or their terms and conditions; **including but not limited to progressively redesigning their content recommendation algorithms to reduce and mitigate risks identified under Article 26(1) (c) to place less emphasis on the retention of attention and more on emphasis of the protection of fundamental human rights and ensuring adequate labelling and notification to users of legal but harmful content on their service;** b) targeted measures aimed at limiting the display of advertisements in association with the service they provide; (c) reinforcing the internal processes or supervision of any of their activities in particular as regards detection of systemic risk; **including but not limited to ensuring that the data sets that inform the detection of risk do so in all relevant languages in which the services operate and do so in a manner that promotes diversity and inclusion and does not breach fundamental rights to freedom form discrimination.** (d) initiating or adjusting cooperation with trusted flaggers in accordance with Article 19; **including but not limited to measures to ensure the sustainability and independence of fact checkers.** e) initiating or adjusting cooperation with other online platforms through the codes of conduct and the

	crisis protocols referred to in Article 35 and 37 respectively **f) The mitigation measures must be designed with the involvement of representatives of the recipients of the service, representatives of groups potentially impacted by their services, independent experts and civil society organisations.**

The adoption of the Code of Practice on Disinformation being developed by the Commission could play a significant part in the success or failure of the DSA in holding a clear line between the removal of illegal content, and the proper procedural management of the risks posed by legal but harmful disinformation content through the risk assessment and mitigation process formed by Articles 26 and 27. **Please see below for our full proposals regarding regulatory backstop for the Code of Practice on Disinformation under Article 35 but its adoption is an obvious measure that should be introduced either through Article 27 (e) or if required through Article 35.**

Mitigation and Article 7 - the prohibition of a general monitoring obligation

Finally we wish to address the question of whether any of the suggested amendments to Article 27 could impinge on protections given to citizen's communications by either Article 7 of the proposed DSA, namely the prohibition of a general monitoring obligation. We do not believe either concern has standing in law because:

1) The kinds of data analyses required to assess and mitigate risks of such content are **specific and proportionate** to the harm disinformation poses to fundamental human rights. They are defined and limited within the article's text and can be further defined through the adoption of the Code of Practice on Disinformation under Articles 27(e) or 35.

2) Such actions are **analogous to due diligence**: they cover specific known harms and are not generalised monitoring. To interpret a prohibition on general monitoring obligations as a "no obligation to monitor even known harms" would effectively free companies from any accountability for content they already curate in their platforms. Surely that illogical result cannot be what is intended by the DSA.

Co-regulatory backstop

With the rapidly changing digital landscape it is absolutely right that the DSA adds a degree of flexibility to the way in which its procedural rules to assess and mitigate systemic risks by drawing in Codes of Practice where those risks are not being fully addressed by Services regulated under the DSA.

As we indicated in our introduction, Avaaz has a particular concern about the effect of disinformation on our societies and would argue that an effective Code of Practice on Disinformation is likely to be needed to address this harm. Disinformation is spread in hugely varying forms, where the line between illegal content, such as hate speech, and legal but harmful, divisive disinformation on an issue such as immigration is not clear. Usually there is a huge grey area between the borders of these two categories.

The DSA answers with its human rights framed processes in Articles 26 and 27 to identify risk and mitigate, and with Article 35 the DSA provides the mechanism by which codes of conduct to refine the ways in which mitigation occur can be given regulatory authority.

We propose specifically that the learnings that we will gain through the Code of Practice on Disinformation are fully integrated into the risk framework set by the DSA through its adoption under Article 35. And suggest the addition of the following paragraph as new paragraph 3:

Current Draft	Suggested amendment
Article 35 [...]	Article 35 [...] 3. **These codes may set out commitments to take specific risk mitigation measures in addition to those on Article 27, as well as a regular reporting framework on any measures taken and their outcomes.**

3. When giving effect to paragraphs 1 and 2, the Commission and the Board shall aim to ensure that the codes of conduct clearly set out their objectives, contain key performance indicators to measure the achievement of those objectives and take due account of the needs and interests of all interested parties, including citizens, at Union level. The Commission and the Board shall also aim to ensure that participants report regularly to the Commission and their respective Digital Service Coordinators of establishment on any measures taken and their outcomes, as measured against the key performance indicators that they contain	~~3~~ **4.** When giving effect to paragraphs 1 and 2, the Commission and the Board shall aim to ensure that the codes of conduct clearly set out their objectives, contain key performance indicators to measure the achievement of those objectives and take due account of the needs and interests of all interested parties, including citizens, at Union level. The Commission and the Board shall also aim to ensure that participants report regularly to the Commission and their respective Digital Service Coordinators of establishment on any measures taken and their outcomes, as measured against the key performance indicators that they contain.
4. When giving effect to paragraphs 1 and 2, the Commission and the Board shall aim to ensure that the codes of conduct clearly set out their objectives, contain key performance indicators to measure the achievement of those objectives and take due account of the needs and interests of all interested parties, including citizens, at Union level. The Commission and the Board shall also aim to ensure that participants report regularly to the Commission and their respective Digital Service Coordinators of establishment on any measures taken and their outcomes, as measured against the key performance indicators that they contain.	~~4~~ **5.** The Commission and the Board shall assess whether the codes of conduct meet the aims specified in paragraphs 1 and 3, and shall regularly monitor and evaluate the achievement of their objectives. They shall publish their conclusions
5. The Board shall regularly monitor and evaluate the achievement of the objectives of the codes of conduct, having regard to the key performance indicators that they may contain	~~5~~ **6.** The Board shall regularly monitor and evaluate the achievement of the objectives of the codes of conduct, having regard to the key performance indicators that they may contain

In conclusion

The inability of social media platforms to systematically address the societal harms they produce – whether during a global pandemic or key elections in the past 12 months alone – confirms the fundamental role that the Digital Services Act can have in determining the future of our digital public space and therefore of our democracies. If the EU wants the Act to be fully effective, it will need to be amended based on the evidence presented in recent research and investigations that shows: first and foremost, the need to keep platforms accountable for the harm created by their own services, and how transparency to users is a fundamental antidote to disinformation and harmful content. It is time for effective, balanced and data-driven regulation. And the DSA is our once in a lifetime opportunity to achieve it.

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[9] revealed dangerous levels of incite ful 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.[10]

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 digital services algorithms, 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

[9] https://secure.avaaz.org/campaign/en/disinfo_hub/

[10] https://secure.avaaz.org/campaign/en/facebook_threat_health/

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 tailored to influence each one of us individually without that process being visible to us. 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 in Digital Services?

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. This is connected to their research on AI inferences about personality type through Facebook 'Likes'. 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 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.

Prohibition on Discrimination

Much has been said about the manner in which algorithmic content accelerators and recommenders are 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.

How are these rights engaged in Digital Services?

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 conceptual framework of consents under GDPR 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 in Digital Services?

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. 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."