Reset—written evidence (FEO0035)

House of Lords Communications and Digital Committee inquiry into Freedom of Expression Online

1. Foreword to response

1.1. A discussion about freedom of expression online often starts with reference to the "trade off" between regulating digital media platforms whose products harm the public and preserving free speech. The argument suggests that it's a zero sum game - that dialing-up public safety means dialing down freedom of speech. This rhetoric is false and undermines admirable attempts by governments such as the UK to intervene in the information marketplace. It is a narrative promoted by the tech companies whose power and resources have allowed this argument to dominate the debate. The truth is that there have always been rules in the digital media marketplace. Some are set by governments -- such as prohibitions on incitement to violence; and most are set by the companies themselves, policing content to comply with social norms, customer expectations, and democratic values. These rules are often ignored. They are applied inconsistently. And they are but rarely explained with respect to how they enable rather than restrict a greater range of public expression.

1.2. Freedom of speech can be preserved while implementing a robust digital regulatory regime. The key lies in altering the architecture of the platforms and the logic of content curation -- that is, how digital media platforms decide what content to serve to consumers. The core of the problem we face today is not primarily a result of malign content appearing on social media. The problem is that the business model of platforms is designed to amplify and spread that content which best captures attention that can be sold to advertisers. That means the recommender or curation algorithms have a bias towards the sensational and outrageous -- which is often synonymous with the malign. Content that once remained at the margins of the marketplace of ideas is systematically dragged into the mainstream and normalized through repetition before a broad audience. This is not about the right of an individual to post on social media; it is about the decision of the platforms to distort the public sphere by amplifying harmful-but-sensational views and sentiments far out of proportion to their actual prevalence in the real world. The first step in harm mitigation should therefore be (re)designing these curation systems so that harmful material stops being algorithmically and automatically spread at scale. People should remain free to state their opinions online (providing they do not break the law in doing so or violate the terms of service of a commercial contract), but those opinions do not have to reach the widest audience in ways that sacrifice the public interest for advertising revenue. It is the difference between freedom of speech and freedom of reach. The former is a

democratic value. The latter is a false claim on the right to manipulate a broken business model.

1.3. Tech companies have demonstrated on many occasions that they have the creativity and tools to apply "harm reduction by design" features to preserve freedom of speech while reducing reach. In such instances, they allow users to share their views while at the same time minimizing the harmful impact. Examples of these features are included in this response.

1.4. In addition, the scale of the threat to freedom of speech online (and the impact of company policies and government rules) cannot be truly understood without greater transparency about how companies design and enforce their policies. At present, private companies decide who can access their services, what they can say and what the repercussions are for crossing self-defined boundaries. There is little to no transparency about how and when they enforce policies which might infringe freedom of speech, nor whether these rules are consistently applied. The companies hold all the information about what freedom of speech looks like online, a situation which must be rectified if policymakers can develop meaningful interventions to protect digital rights.

1.5. The market power and control over national information distribution held by these platforms exacerbates this problem. In a world in which a few Big Tech companies control virtually all of what we see on digital media, it is unacceptable that the public has no visibility into how they make content moderation decisions. As commercial providers of information, they are and should be free to set rules for the use of their products. But their market dominance demands that these practices are transparent. Consider that we do not subject *The Guardian* or *The Telegraph* to government oversight of the criteria for why they do or do not publish certain views on their editorial pages. But if they were the only two news distributors in the nation, the public could and should demand to know more about how corporate practices align with the needs of a democratic public sphere.

1.6. The UK has made some significant advances in developing a regulatory framework which preserves freedoms while minimising harm. The duty of care at the heart of HMG's Online Harms agenda offers the most promise, since it requires companies to build in harm reduction measures by design rather than delete harmful content once it has been published (unless it is clearly illegal). This duty of care should be extended to cover democratic harms, which the Full Government Response excluded from the online harms agenda. Recent events in the United States have shown the havoc democratic harms such as electoral disinformation can reap in the offline world. Rather than leave the companies to self-regulate free speech in this category of harm, the Government must revisit how regulation can be reinforced to avoid the chaos witnessed in recent weeks.

2. Response to questions

Q1. Is freedom of expression under threat online? If so, how does this impact individuals differently, and why? Are there differences between exercising the freedom of expression online versus offline?

2.1 Freedom of expression is under threat online, but not just in the ways the loudest, coarsest voices on the internet lead us to believe. As the preface to this inquiry notes, there have been a number of high-profile figures "de-platformed" by Facebook and Twitter for comments they make in breach of platform T&Cs. The recent blocking of President Trump is by far the most prominent example -- though notably the circumstances in which it occurred are historically extraordinary. Rightly, critics berate the platforms for inconsistent application of their policies and lack of transparency about why such decisions are made. But these individuals are often media-savvy, well-followed celebrities with the power and resources to make their voices heard through various channels. Their status affords them the freedom to express themselves and to continue to be heard by a wide audience. Meanwhile, a large number of people are being silenced online without recourse.

2.2 We should pay special attention to the problem known as "censorship through noise" -- literally online harassment to such a degree that it chills speech and infringes rights. Women, BAME and LGBTQ communities are just some of the groups whose freedom to engage online is being quietly eroded by the torrent of abuse and harassment they witness on the internet. In a recent survey, 78% of British women who expressed an opinion felt Twitter was not a place they can share their opinion without receiving violence or abuse.[1] A separate survey by the charity Plan International concluded that women and girls across the world are "opting out of expressing themselves and their opinions for fear of retribution, and sometimes removing themselves from these platforms altogether".[2] This threat of online abuse extends to women from all walks of life, be they politicians, journalists or young girls. Last year, a group of over 100 female legislators from all over the world wrote to Facebook asking the company to stop the spread of gendered disinformation and misogynistic attacks against women leaders. In the letter, the politicians called for Facebook to "safeguard online spaces to allow all voices to be heard without manipulation, harassment, or intimidation".[3] This is the less visible, more insidious censoring of people who are afraid to express themselves online for fear of abuse of persecution. We must look carefully at our principles when we assess whether reducing the reach of a speaker engaged in harassment is justified in order to safeguard the rights of harassed groups to speak and for a

---

1 Toxic Twitter - A Toxic Place for Women

2 Abuse and harassment driving girls off Facebook, Instagram and Twitter, Plan International, Oct 2020

3 Letter to Facebook, August 2020

broader public to hear them. Free speech is not merely a negative individual right; it is a positive, collective right.

2.3 While different demographics are silenced in different ways in the online world, they are all victims of the same structural problem - the lack of transparency about how technology platforms make decisions about who can stay on their platforms or why. While some users are deplatformed without notice or an opportunity to appeal, others are forced to self-silence rather than wait for those harassing them to be rebuked by the platforms. This lack of information makes it impossible for policymakers to devise and measure meaningful interventions to preserve free speech online. It also means the users of these platforms have little sense of where the boundaries are - when they might be blocked from a site, and for what reasons. Improving transparency is crucial to mitigating freedom of speech infringements, and to ensuring companies' policies are implemented fairly and consistently.

Q2. How should good digital citizenship be promoted? How can education help?

2. 4 No response

Q3. Is online user-generated content covered adequately by existing law and, if so, is the law adequately enforced? Should 'lawful but harmful' online content also be regulated?

2.5. Recent initiatives by the UK government to replace the current system of self-regulation by companies with a balanced, consistent application of T&Cs are admirable and, in some instances, world-leading. The Age Appropriate Design Code is an exemplary policy initiative to protect vulnerable groups online. The upcoming Online Safety Bill looks to have similar merits and should be praised for promoting a duty of care, which requires in-built prevention of harm rather than a "takedown" approach that infringes on freedom of speech and expression.

2.6. However, in addressing legal but harmful content the Full Government Response to the Online Harms White Paper has excluded "democratic harms'' from scope. This is despite calls from many parliamentarians and civil society groups for such harms to be included. Sadly, events surrounding the US election were the perfect case study of how democratic harms can have troubling implications for broader society, even resulting in loss of life. Electoral disinformation, which was generated on an astounding scale during the Presidential election and for months after, can trigger a whole range of harms - lawful and unlawful. Such harms should absolutely be included in the Online Safety Bill, as they are in the EU's Digital Services Act. As present, the proposals set out in the Full Government Response do not go far enough to make the UK the safest place to be online.

2.7 To address the issue of enforcement - specifically ineffective and unenforced terms and conditions - any regulator of online harms should include in its monitoring and compliance toolkit the power to inspect algorithmic systems -- from the data they use to train the software through to the impact on particular social groups.[4] In order to regulate online harms, regulators such as Ofcom need to understand how these powerful lines of code disseminate harmful content. As algorithms are designed and deployed at unprecedented scale and speed, there is a pressing need for regulators to keep pace with technological development; they must establish the systems, powers, and capabilities to scrutinise algorithms and their impact. Having this authority to audit the algorithms amplifying harmful material on social media platforms is by far the best way for regulators to understand and monitor companies' policies, processes, and data. A paper from November 2020 by the Ada Lovelace Institute and Rest on the role of algorithmic inspection and audit covers the issue in full.[5]

Q4. Should online platforms be under a legal duty to protect freedom of expression?

2.8. Online platforms should be under a legal duty to reduce the amplification of material which causes harm, including democratic harms. This is at the heart of the duty of care approach set out in the Online Harms White Paper. Such an approach, if properly implemented, preserves freedom of expression by ensuring that legal content remains online while mitigating the harmful impact. In that way, the duty of care is a legal duty to protect freedom of expression.

Q5. What model of legal liability for content is most appropriate for online platforms?

2.9 No answer

Q6. To what extent should users be allowed anonymity online?

2.10 No answer

Q7. How can technology be used to help protect the freedom of expression?

2.11 In response to deluge of misinformation circulating on social media in recent months, tech companies implemented a number of harm reduction measures which protect freedom of speech and demonstrate what harm reduction by design looks like in practice. Some examples are below.

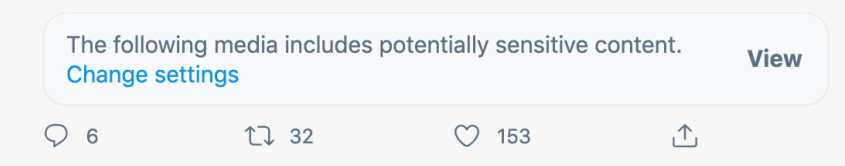
---

4 Algorithm Inspection and Regulatory Access. Demos, doteveryone, Global Partners Digital, Institute for Strategic Dialogue, Open Rights Group. April 2020.

5 Algorithms in social media: realistic routes to regulatory inspection, November 2020

Twitter's warning messages - sensitive media

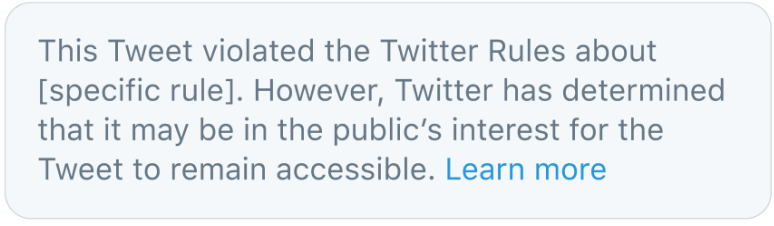
2.12 Twitter has some of the most prominent examples of harm reduction by design features, notably the warning messages it applies to certain content. These messages are overlaid on specific Tweets, warning users about the nature of the content in the Tweet and requiring them to click through before they can view it. The Tweets stay on the site - content is not removed.



2.13 Before the summer of 2020, these messages were only applied to content that has been marked (either by the person Tweeting it or following reports by other users) as "sensitive", such as media that included adult content, excessive gore or violence. This reduced the risk of users inadvertently witnessing content they might find harmful or distressing, but allowed users who did want to find such content to access it. Users can choose whether to turn this feature on/off, so they don't have to click through to view sensitive content.

Twitter's warning messages - public exemption policy media

2.15 In June 2020, Twitter applied for the first time its "public exemption policy". The policy states that when a Tweet contains harmful content but is deemed to be in the public interest, the Tweet will be placed behind a notice. Such content would include harassment or hateful conduct, content which is in breach of Twitter's T&Cs and for the majority of users would have to be taken down. Instead, in such instances, the notice would be applied which still "allows people to click through to see the Tweet" but "limits the ability to engage with the Tweet through likes, Retweets, or sharing on Twitter, and makes sure the Tweet isn't algorithmically recommended by Twitter". The exception only applies to elected or government officials with over 100,000 followers, and aims to "limit the Tweet's reach while maintaining the public's ability to view and discuss it".



2.16 Limiting a message's reach because it is deemed to include harmful information, rather than removing it entirely, is an artful way of handling "legal but harmful" content. It tactfully navigates freedom of speech concerns, allowing information to remain in the public domain while reducing the level of public exposure and engagement. Such checks and balances play a critical role in slowing the spread of harmful content, and are particularly crucial at a time when false information is proliferating at an unprecedented pace. Freedom of speech does not infer maximum freedom of reach, and reducing the reach of content in this manner is a nuanced, proportionate, and human rights friendly option. The Online Safety Bill should be incentivising companies to change their systems in this way to reduce harm.

Q8. How do the design and norms of platforms influence the freedom of expression? How can platforms create environments that reduce the propensity for online harms?

2.17 See above answer.

Q9. How could the transparency of algorithms used to censor or promote content, and the training and accountability of their creators, be improved? Should regulators play a role?

2.18 Regulators should absolutely have the powers to audit algorithms. As discussed in this response, at present there is an inappropriate lack of transparency about how companies determine what information is algorithmically promoted to individuals. Companies should be required to share certain data about the design of their platforms with regulators, governments and academia. Without access to the data and AI systems that guide information flows in these markets, there is no obvious way to make good policy that will be adaptive and durable as the industry evolves. There need to be new systems for transparency and auditing of algorithmic design and decision making, giving regulators the powers and tools to inspect these powerful lines of code.

2.19 The Online Harms White Paper identified this problem, stating that regulators should be able to "require additional information, including about the impact of the algorithms" and to "request explanations about the way algorithms operate". This does not go far enough. Regulators need to have the tools and powers to test the operation of algorithms and to undertake inspections themselves. At present, there is a massive asymmetry of information. The harms are easily observed as specific incidents, and they do in fact appear to form a pattern. But the companies that hold the data that could verify these patterns and measure their scope hold all the data, and they do not make it available for independent review under any circumstances. This lid is kept tightly shut. Without access, regulators are forced to rely on the companies to police themselves through ineffective codes of conduct. This is extraordinary. We have an industry operating in

markets with clear externalities that cause public harms. The companies have all the data and tools needed to track, measure and evaluate these harms - indeed these tools are a core part of their business. But they make none of these available to public oversight, even as they avoid all but the most basic interventions to protect the public from harm.

2.20 There is precedent in the UK for a regulator to have such powers of oversight. The Information Commissioner's Office (ICO) has licence to undertake consensual audits to assess how data controllers or processors are complying with good practice in the processing of personal data.[6] Should the company not agree to a consensual audit, the ICO can seek a warrant to enter, search, inspect, examine and operate any equipment in order to determine whether a company is complying with the Data Protection Act.[7] Similarly, the Investigatory Powers Commissioner's Office (IPCO) has powers[8] to conduct investigations, inspections and audits as the Commissioner considers appropriate for the purpose of the Commissioner's functions, including access to apparatus, systems or other facilities or services.[9]

2.21 Ofcom will need a similar ability to carry out an algorithm inspection with the consent of the company; or if the company doesn't provide consent, and there are reasonable grounds to suspect they are failing to comply with requirements, to use compulsory audit powers. The resource to carry out these investigations could sit within the regulator, but they could also have the power to instruct independent experts to undertake an audit on their behalf. This would help ensure that the correct expertise is acquired for the work as is needed. This would mirror the Financial Conduct Authority's power to require reports from third parties; what they dub "skilled persons reviews".[10]

2.22 In addition, as recommended by the Centre for Data Ethics and Innovation, academics should be able to access certain datasets when conducting research into issues of public interest.[11] Efforts in this area are underway[12], but they have been challenging to establish, are limited in scope and are yet to prove themselves. While the online harm regulator will be able to "encourage" companies to give researchers access to data, its powers will need to go beyond mere encouragement. The power of these datasets should, in certain circumstances, be available to serve the wider public good.

---

6 s129, Part 5, Data Protection Act 2018
7 Schedule 15, Data Protection Act 2018
8 s235(1), Chapter 1, Part 8, Investigatory Powers Act 2016
9 s235 (4), Chapter 1, Part 8, Investigatory Powers Act 2016
10 Skilled person reviews, Financial Conduct Authority
11 Review of online targeting, Centre for Data Ethics, February 2020
12 https://socialscience.one/

Q10. How can content moderation systems be improved? Are users of online platforms sufficiently able to appeal moderation decisions with which they disagree? What role should regulators play?

2.23 No answer

Q11. To what extent would strengthening competition regulation of dominant online platforms help to make them more responsive to their users' views about content and its moderation?

2.24 Events in the US over the past week have demonstrated what happens when tech giants collectively apply their power in the freedom of speech arena. The blocking of President Trump from Facebook and Twitter has raised questions about editorial control by the platforms, although by inciting violence his comments were squarely in breach of platform rules. Where competition comes into play are the moves by Apple, Google and AWS to effectively remove Parler, a challenger social media platform, from the internet. Market dominance enables a small number of companies to decide not only what content people see on the internet, but also to crush competitors overnight. Such moves to ban services by the owners of market bottlenecks (e.g. mobile app stores) demonstrate why the work of regulators like the Competition and Markets Authority is so relevant and pressing. At the very least, there should be transparency rules applied to these kinds of decisions -- even if there is good evidence that the bans are warranted due to illegal activity (as in the case of Parler). The CMA and its new Digital Markets Unit should be suitably resourced to investigate and regulate competition issues in information markets, and have the powers to intervene where necessary.

Q12. Are there examples of successful public policy on freedom of expression online in other countries from which the UK could learn? What scope is there for further international collaboration?

2.25 Regulating the internet and preserving freedom of expression are global issues which should be tackled at an international level. In recent years, the UK has led the charge in delivering a robust digital regulatory regime. However, it has lost some ground to the EU with the publication of the Digital Services Act and risks conceding further in its efforts to make the UK the safest place to be online. The UK has the opportunity to regain this ground and to lead the agenda via its upcoming presidency of the G7, and should look to work with colleagues across the G20 and Commonwealth to build global alliances and standards. In addition, as the UK seeks more trade deals it must ensure that agreements do not undermine efforts to protect freedom of expression online nor to develop tough digital regulations.

About Reset

Reset (www.reset.tech) was launched in March 2020 by Luminate in partnership with the Sandler Foundation. Reset seeks to improve the way in which digital information markets are governed, regulated and ultimately how they serve the public. We will do this through new public policy across a variety of areas – including data privacy, competition, elections, content moderation, security, taxation and education.

To achieve our mission, we make contracts and grants to accelerate activity in countries where specific opportunities for change arise. We hope to develop and support a network of partners that will inform the public and advocate for policy change. We are already working with a wide variety of organizations in government, philanthropy, civil society, industry and academia.

15 January 2021