Reset.

Video sharing platform regulation

Part B. Questions for all stakeholders

Question 19: What examples are there of effective use and implementation of any of the measures listed in article 28(b)(3) the AVMSD 2018?

RESPONSE

Terms and conditions and flagging and reporting mechanisms - While not a VSP in UK jurisdiction, Twitter has delivered some instructive examples of effective measures a VSP can take to reduce harm in recent months.

First, in March 2020, and in response to the disinformation flooding social media platforms during the early days of the pandemic, Twitter broadened its definition of harm to include denying health authority recommendations, fake COVID-19 treatments, creating panic based on fake claims, impersonating government health officials, and fake claims about immunity or susceptibility for certain groups. This expanded definition set a precedent in recognising the consequences of “legal but harmful” content.

In addition, in May 2020, Twitter applied a public interest notice to two Tweets by President Trump, the first time Twitter has applied the notice to any Tweets since it developed its “public interest exceptions” policy in June 2019. The policy states that while Twitter usually removes Tweets which violate its rules, in some instances, where the Tweet is deemed to be in the public interest, the Tweet will instead be left up but placed behind a notice. The notice 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”.

This intervention by Twitter demonstrated, in a very public way, how an online harms reduction strategy can be implemented practically through a feature or design change to the product. 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. The rights associated with freedom of speech do not extend to the “freedom of reach” that may come through the algorithmic amplification of content by recommendation and curation systems. Reducing the reach of content by blunting the application of recommendation algorithms and reducing visibility through design changes represent nuanced, proportionate, and human rights friendly options for addressing harmful content.

In the context of the upcoming Online Harms Bill, Twitter has effectively shown that harms reduction by design is not only possible, but that it is already live, part of existing company policy and readily deployable. This is reassuring, particularly in the current climate where people increasingly rely on elected and government officials for trustworthy information. Carnegie UK Trust has undertaken a detailed human rights analysis, focussing on freedom of expression, which sets out further detail of how focussing regulation on system changes rather than regulating individual content is the right approach to take. The Online Harms Bill should be incentivising companies to change their systems in this way to reduce harm.

Question 20: What examples are there of measures which have fallen short of expectations regarding users’ protection and why?

RESPONSE

Terms and conditions

Despite the response to question 19, in general VSP terms and conditions by themselves are ineffective at protecting users. This is in part because they are voluntary standards, and inconsistently enforced by the platforms. Each platform develops their own policies with limited statutory or regulatory obligations to consider, and releases transparency reports at their discretion. The platforms “hold all the cards: they draw up their terms of use, decide to what extent to be bound by them, modify them as necessary without any public formalities.”[1] This creates tremendous information asymmetry; public authorities and civil society cannot access the evidence required to analyse harms related to the platform, or to assess the efficacy of platforms’ policies and interventions.[2] Relying exclusively on T&Cs as a mechanism to protect users does not work.

Moreover, there is the question of verification. Repeatedly throughout the last several years, we have witnessed a recurring pattern with respect to digital platforms and harmful content reduction. The cycle begins with users spotting and reporting harmful content. The response of the company is opaque. When this kind of content goes viral, it then draws media attention.

[1] ARCEP cited in Review of online targeting: Final report and recommendations. Center for Data Ethics and Innovation. May 2020.

[2] Review of online targeting: Final report and recommendations. Center for Data Ethics and Innovation. May 2020.

The companies publicly pledge to address the problem and to do a better job of enforcing their own terms and conditions. Months later, there are still many examples of the harmful content online, and yet the platforms claim that they have largely solved the problem -- often citing a figure like 90% or better to show the efficacy of their performance. Effectively, we have no means to verify if this is accurate or not. The public is left wondering whether the repeated examples of harmful content are indeed anomalous or if the companies have not actually achieved a high standard of harm reduction. Such a policy of "trust-but-don't-verify" would be unacceptable in any other industry that has a similarly broad effect upon the public, e.g. pharmaceutical, financial services, automotive or chemical. It calls into question whether terms of service meaningfully improve consumer outcomes or simply represent a policy for lip service alone.

For example, research has extensively demonstrated how foreign and domestic actors have violated the Terms of Services of most platforms to spread health-related or political disinformation:

- University of Oxford: Russia's IRA spread false information designed to create outrage about Black Lives Matter and deepen social division in the US.
- Nature: Game theory analysis has shown how a few bots with extreme political views, carefully placed within a network of real people, can have a disproportionate effect within current social media systems. Studies demonstrate how an extremist minority political group can have undue influence using such bots—for example, reversing a 2:1 voter difference to win a majority of the votes.
- Avaaz: Two separate, quantitative analyses demonstrate how COVID-related disinformation is widespread on Facebook and remains largely unchecked by terms of service restrictions.[3]

Platforms have addressed this in part by outlawing what they have dubbed "coordinated inauthentic behaviour" or other forms of misinformation. However, the question remains whether our democratic societies should leave it to the platforms to determine which accounts are "inauthentic" and/or which set of activities can be seen as "coordinated". Platforms are also not very good at finding coordinated inauthentic behaviour in all the languages that they operate in. Currently we don't know how exactly platforms decide if coordination is the result of a grassroots movement or a carefully planned "influence operation". And once again, there is no mechanism for oversight or verification of these internal company processes -- despite their importance in shielding the public from harm.

Handling and resolving users' complaints

In recent months, digital media companies have attempted to tackle the surge in false information circulating on their platforms. Ofcom's own polling showed that exposure to

[3] See: How Facebook Can Flatten the Curve of the Coronavirus Infodemic; and Facebook's Algorithm, a Major Threat to Public Health.

disinformation is worryingly high, with 50% of respondents encountering false or misleading information on a weekly basis in the early period of the pandemic.[4] Notably, this extraordinary level of exposure is NOT declining week after week -- despite much publicised attempts by tech companies to reduce disinformation.

Disinformation circulated so rapidly during the early weeks of the pandemic that firms struggled to catch up. Content promoting conspiracy theories harmful to public health is being viewed millions of times before it can be removed or flagged as untrustworthy, only to reappear elsewhere in the digital realm.[5] NewsGuard's recent data highlights how certain Facebook Pages are being used as "super-spreaders" of Covid-19 disinformation, amplifying fake news across the platform: 36 European Facebook Pages alone reach 13,223,446 Facebook users.[6] A recent report by Avaaz found that only 16% of all health misinformation had a warning label, and that, despite being fact-checked, 84% of misleading articles and posts that Avaaz sampled remained online without warnings. [7]

The repercussions of not removing content quickly are real. One paramedic for NHS 111 says he spends 10 minutes on average with each patient correcting misconceptions they have picked up about Covid-19. He pointed out the information asymmetry saying, the conspiracy 'Plandemic' video on YouTube got 40 million views in 48 hours which was like spreading misinformation to 25,000 people in 10 minutes.[8]

This relentless flow of information, and disinformation, is powered by machine-learning algorithms and automation. These algorithms serve the business model by maximising users' time spent on the platform - the number of clicks, shares, and pageviews that drive advertising sales. In a kind of pincer movement on the user, the social media platforms curate noncommercial (i.e. organic) information flows to optimise for engagement even as they sell targeted advertising designed to maximise influence over preferences. The unimaginable size of the training data, the world-leading investment in AI experts, and the scale of the testing interface delivers efficient results and ever-increasing quarterly returns for the Silicon Valley giants. As we spend more time online during the Covid-19 lockdown[9], these algorithms are being fed with even more data. And, even more worryingly, more people are seeking out reliable information about public health and getting back disinformation, conspiracy, and anti-vaccination propaganda instead.

---

[4] Covid-19 news and information: consumption, attitudes and behaviour, Ofcom, May 2020
[5] How the 'Plandemic' Movie and Its Falsehoods Spread Widely Online, New York Times, 20 May 2020
[6] Facebook 'Super-spreaders': Europe, NewsGuard, 5 May 2020
[7] https://secure.avaaz.org/campaign/en/facebook_threat_health/, August 2020
[8] Paramedic Thomas Knowles testimony to the Select Committee - https://publications.parliament.uk/pa/cm5801/cmselect/cmcumeds/234/23404.htm#footnote-230
[9] Locked-down households using internet for 41 hours a week, USwitch, 5 May 2020

Question 21: What indicators of potential harm should Ofcom be aware of as part of its ongoing monitoring and compliance activities on VSP services?

RESPONSE

Legal but harmful

Harms come in many forms, and Ofcom should take an expansive definition of harm to include legal but harmful material. While partially within scope for the VSP regulation, legal but harmful material could - and, in our view, *must* - be included in the upcoming online harms legislation. Without its inclusion the ills of recent months - rampant disinformation, fraud and profiteering, as well as the exploitation of vulnerable groups - will continue to proliferate unabated (despite the efforts of the companies at marginal self-regulation). The example of public health disinformation is a demonstrative example of why harmful-but-not-illegal content must be addressed. Almost none of the disinformation and conspiracy about COVID-19 or vaccinations is illegal on its face. And yet, it has an obviously negative impact on the public, particularly when it is amplified through social media recommendation algorithms.

Algorithmic inspection - enforcing terms and conditions

To address the issue of ineffective and unenforced terms and conditions, Ofcom 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.[10] Self-regulation ignores the misalignment[11] of incentives between social media platforms – who prioritise targeted advertising at scale, and have designed their infrastructure accordingly – and public policy objectives.[12] Ultimately, the tools used to amplify harms on the platforms are the same used by legitimate businesses; COVID-19 misinformation on the platforms still runs rampant because content promotion and amplification systems that underpin the platforms' business models are largely left untouched by voluntary actions.[13] Indeed, we can hardly expect companies to voluntarily embrace practices that undercut their business models. Social media platforms are built on amplifying sensational content to attract attention that they then sell to advertisers

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

[11] Review of online targeting: Final report and recommendations. Center for Data Ethics and Innovation. May 2020.

[12] It's the (Democracy-Poisoning) Golden Age of Free Speech. Z. Tufekci. *Wired*. 16 January 2018. See also Facebook co-founder Chris Hughes' statement: ""Facebook's business model is built on capturing as much of our attention as possible to encourage people to create and share more information about who they are and who they want to be."

[13] The First 100 Days: Coronavirus and Crisis Management on Social Media Platforms. C. Colliver and J. King, *Institute for Strategic Dialogue.* 2020.

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 VSPs is by far the best way for Ofcom to understand and monitor companies' policies, processes, and data.

A regulatory inspection of algorithmic systems, often referred to as an audit of algorithmic systems, is a broad assessment approach of an algorithmic system's compliance with regulation.[14] In the case of social media platforms and misinformation, this activity is forward-looking; the regulation in question is not yet in place, although it is under discussion through the Online Harms Bill in the UK and the Digital Services Act in the EU.

Any regulatory policy of algorithm inspection would be the first of its kind – no regulator has conducted one before. Until now, audits and inspections have been conducted by technology companies as internal exercises, or by external researchers with limited access to technology companies' data, policies, and processes.

Question 22: The AVMSD 2018 requires VSPs to take appropriate measures to protect minors from content which 'may impair their physical, mental or moral development'. Which types of content do you consider relevant under this? Which measures do you consider most appropriate to protect minors?

RESPONSE

On this issue, we would direct Ofcom to the response from 5Rights Foundation.

Question 23: What challenges might VSP providers face in the practical and proportionate adoption of measures that Ofcom should be aware of?

RESPONSE

VSPs pride themselves on their ability to innovate at pace and at scale. VSPs are some of the world's largest and most valuable companies - they will be well-resourced to adapt to these new measures. They have also been calling for greater regulatory oversight, and so should welcome efforts to reduce the amount of harmful content circulating on their platforms.[15] The VSP Regulation does not ask too much of VSPs - it mostly expects them to enforce terms and

[14] Examining the Black Box: Tools for Assessing Algorithmic Systems. Ada Lovelace Institute, DataKind UK. 2020

[15] BBC Radio 4, *Today Programme,* 24 June 2020

conditions they already have in place. Adopting these measures should not be too arduous for VSPs. To the extent reasonable, OfCom can adopt a principle of proportionality when applying new policies that recognize the differences in scale, resources and public impact represented by differently sized companies.

The bigger issue is how any regulator prepares and resources itself for implementing regulation effectively, especially when the subjects being regulated are some of the world's most valuable companies. Algorithmic auditing, for example, requires a multidisciplinary skill set. While the regulator should have some skills in-house, it will need the ability to access and instruct third-party expertise. This could be through powers similar to those of the UK's Financial Conduct Authority, who can require reports from third parties, or through a new field of registered auditors. Alternatively, the regulator could give independent experts secure access to platform data to undertake audits on its behalf.

As recommended by the Centre for Data Ethics and Innovation[16], academics should be able to access certain datasets when studying issues of public interest. The regulator should have the powers to mandate this access, especially on issues such as disinformation, where independent research will be crucial to developing future public policy.

Question 24: How should VSPs balance their users' rights to freedom of expression, and what metrics should they use to monitor this? What role do you see for a regulator?

RESPONSE

At the heart of the upcoming Online Harms Bill is a duty of care - a systemic approach to regulation which does not rely on removing content. Carnegie UK Trust has undertaken a detailed human rights analysis, focussing on freedom of expression, which sets out further detail of how focussing regulation on system changes rather than regulating individual content is the right approach to take.

As recent efforts by technology companies have shown, taking down content doesn't work. Harmful information spreads at such speed, the platforms struggle to keep up. A systemic approach which reduces the amplification of harmful material, without infringing on freedom of expression, is a much better and effective way forward.

---

[16] Review of online targeting: Final report and recommendations. Center for Data Ethics and Innovation. May 2020.

Question 25: How should VSPs provide for an out of court redress mechanism for the impartial settlement of disputes between users and VSP providers? (see paragraph 2.32 and article 28(b)(7) in annex 5).

NO RESPONSE

Question 26: How might Ofcom best support VSPs to continue to innovate to keep users safe?

RESPONSE

Ofcom should encourage innovation by setting high regulatory standards. As mentioned above in our answer to q23, VSPs are some of the world's most innovative companies. They should embrace the challenge to keep users safe. The current model of self-regulation does not push companies far enough - they need, and indeed have called for, external forces to drive them forward. Consider cyber security, a flourishing sector of the economy which emerged almost entirely in response to government regulation and intervention. Governments around the world have set high standards for compliance with cyber security regulations, which has driven innovation in incumbent companies and sparked the establishment of many new ones. The same can be true for VSPs and social media companies in tackling online harms. Governments and regulators should aim for the gold standard, and VSPs should embrace the challenge.

Question 27: How can Ofcom best support businesses to comply with the new requirements?

NO RESPONSE

Question 28: Do you have any views on the set of principles set out in paragraph 2.49 (protection and assurance, freedom of expression, adaptability over time, transparency, robust enforcement, independence and proportionality), and balancing the tensions that may sometimes occur between them?

RESPONSE

Together, these principles allow Ofcom to deliver a robust regulatory regime while encouraging creativity, innovation and a positive user experience. The most sensitive balance to be struck is between robust enforcement and freedom of expression. The public policy agenda must be implemented in ways that preserve freedom of expression, privacy and security. A content-focused approach which relies on removing harmful content runs the risk of infringing expression rights. A more systemic approach, as should underpin the Online Harms Bill, maintains a fair balance - allowing users to say and share legal content, but reducing the amplification of harmful content when it is circulating. Ofcom should draw on the duty of care principle when delivering and devising future harm reduction measures.

As discussed in this response, transparency is only meaningful if the regulator has the power and resources to get to the heart of what is driving harms online - automation, algorithms and data. 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. None of the issues at the centre of this debate can be adequately addressed without this auditing function: democratic election integrity, child online safety, anti-competitive practices, consumer fraud and abuse, harassment and hate speech, and much more. 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.

The presence of harmful content in the public sphere is nothing new. What is new is the application of AI to the business of information distribution and targeting that enables forms of artificial amplification of harmful content, increases the chance of high frequency exposure to extreme views, and opens up media channels that double as interpersonal communications networks to organised exploitation. It is these processes that must be examined and regulated to comply with standards consistent with democratic values.

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.

There is precedent in the UK for a regulator to have such powers of regulatory access and 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.[17] 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.[18] Similarly, the Investigatory Powers Commissioner's Office (IPCO) has powers[19] 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.[20]

Any future regulator of online harms 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".[21]

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.[22] Efforts in this area are underway[23], 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. A joint paper prepared by Demos, doteveryone, Global Partners Digital, Institute for Strategic Dialogue and Open Rights Group, coordinated by Digital Action, provides more detailed recommendations.[24]

The following table was produced as the output of a jointly co-hosted workshop between Reset and the Ada Lovelace Institute in August 2020 which sets out the methods and types of access that algorithm inspection could provide. It will be included in an upcoming paper soon to be published by Reset and the Ada Lovelace Institute which will make additional recommendations on algorithm inspection in the context of social media platforms.

[17] s129, Part 5, Data Protection Act 2018
[18] Schedule 15, Data Protection Act 2018
[19] s235(1), Chapter 1, Part 8, Investigatory Powers Act 2016
[20] s235 (4), Chapter 1, Part 8, Investigatory Powers Act 2016
[21] Skilled person reviews, Financial Conduct Authority
[22] Review of online targeting, Centre for Data Ethics, February 2020
[23] https://socialscience.one/
[24] Algorithmic Inspection and Regulatory Access, Demos, doteveryone, Global Partners Digital, ISD, Open Rights Group and Digital Action, May 2020

Method	Examples	Benefits	Challenges
Documentation	Policy documentation, including definitions of misinformation or harmful content, related platform rules, and actions and reasoning behind them	Provides evidence of the company's (claimed) expected behaviour Enables initial scrutiny of policy stance	Without details of company processes and systems, risk of being a high-level understanding of policy intent (and not of realities on the platform).
	Process documentation, including instructions given to manual content moderators	Provides evidence of the company's (claimed) expected behaviour Enables initial scrutiny of process design	If made public, risks making it easier to 'game' moderation systems.
	Technical system documentation, including: - tools used to identify and moderate information - content recommendation and sharing systems	Provides evidence of the company's (claimed) expected behaviour Enables initial scrutiny of technical design	If made public, risks making it easier to 'game' moderation systems.
Self-reported metrics	Self-reported metrics on misinformation and harmful content, such as: - Model performance for recommender and moderation systems (including false positives and false negatives) - Commercial data for promoted content that's later moderated. - Engagement metrics for content that's later moderated	Provides evidence of the extent to which company believes it is meeting standards	Lacks independent verification Platforms can selectively choose what to report.
Interviews	Interviews with staff beyond the typical policy and legal teams who interface with regulators, such as:	Direct access to those who design and implement systems will more quickly reveal the	The power dynamic of employer-employee relationship may pressure selected interviewees.

	- Technical staff on product teams focused on moderation and recommendation software (product managers, engineers, data scientists) - Moderation teams implementing policies	principles underpinning the system, and design and engineering decisions and trade-offs.	
Dataset provision	Datasets shared with inspectors could include samples of moderated and unmoderated content and/or training data to develop moderation or recommendation models	Enables independent scrutiny of system, and provides inputs and outputs to verify function and impact	Datasets provide a snapshot of a single point in time - they may become out of date as user behaviour or system algorithms change Datasets may be selective Privacy concerns for users
API access	Access to new or extended APIs for an inspector, such as access to live platform data.	Enables real time/rolling scrutiny of a system's inputs and outputs to verify function and impact	Ongoing access must be agreed upon. Companies could manipulate data available through the API.
Code access	Access to code that underpins moderation or recommendation systems	Allows interrogation of algorithms and verification of system function	Code changes over time; access would need to be ongoing to be meaningful Security threat of ongoing access to systems Privacy concerns for users Understanding the code would require technical expertise (which may vary by platform). This would likely be slow and would benefit from support of engineers working at the social media platform Concerns about intellectual property
Inspector-set test results	A test or dataset for companies to run on their	Access to information and systems that are	Results are not independently verifiable;

	platforms (or for the inspector to run through a private API), in order to collect test results.	not public, without direct access to systems	concerns raised about reliability. Hard to set universal tests for different platforms due to different content formats or processes

---

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.