Your response

Questions for all stakeholders	Your response
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? **The measures are terms and conditions, flagging and reporting mechanisms, age verification systems, rating systems, parental control systems, easy-to-access complaints functions, and the provision of media literacy measures and tools. Please provide evidence and specific examples to support your answer.**	Confidential? – NO. Much of the AVMSD is centred around two main strategies for platforms to moderate online content: banning and using flags. Yet the range of options open to platforms for content moderation is, in practice, far greater. We have identified three broad categories of interventions, which are *already* being used by platforms. They introduce varying levels of friction (from limits on paying for content to be promoted through to fully banning users) in order to disrupt and minimize the spread of online hate. At present, there is a lack of research and critical civic discourse about the impact of these strategies, so we advise (1) the regulator could drive forward research in this area, supporting better theoretical, technical, and ethical understanding of each intervention, and (2) the regulator could direct platforms to be more transparent about the different interventions they use and how this impacts users' activity, such as by providing estimates of the numbers of users and percentage of content which is affected. **Search constraints** 1. Stopping content from being promoted – platforms stop accounts from paying to have their content reach new audiences. YouTube and Facebook both do this. 2. Making content unsearchable – platforms remove certain accounts from their search algorithm. YouTube uses this, such as with the far right figure Tommy Robinson. 3. Making content less visible in searches – platforms 'downvote' some content in their algorithms. Facebook uses this approach for potentially harmful content. 4. Constraining how many times content is shared – content can only be shared so many times before it much be reposted, limiting the spread of harmful 'viral' content. Whatsapp has used this, and Facebook has recently started doing so too. **Viewing constraints**

1. Attaching a ‘warning’ to content – content is flagged as being harmful, toxic or otherwise contentious. Twitter often uses this approach. Many platforms have implemented fact checking flags for content relating to COVID-19 [1].
2. Requesting explicit consent from viewers – content can only be viewed if it has been clicked on. Twitter often uses this approach.
3. Showing content with a competing viewpoint – to our knowledge, this is not currently used by platforms. However, it is a promising way of breaking people out of filter bubbles.

Hosting constraints

1. De-monetising content – accounts can no longer make money from their posts. This is more relevant for platforms with ‘creators’, such as YouTube and Snapchat.
2. Shadow banning – users are banned without knowing that they are banned. They believe they are still posting ‘live’ messages, but they are not seen by anyone on the platform. Reddit is known to do this.
3. Removing content (temporarily) – nearly all platforms suspend content. This is often whilst they make a full decision about it. See the recent case of Wiley on Twitter for an example [2].
4. Removing content (permanently) – content is taken down. Sometimes a notice is left up to make other users aware that the content has been removed.
5. Suspending users – users are temporarily banned. The severity often escalates, with suspension periods increasing the more frequent they are.
6. Banning users – users are banned permanently, and are usually not allowed to create other accounts. This is the most severe option and is usually only used for spam- and bot- accounts and repeated offenders of Terms of Service.

The AVMSD proposes media literacy as a way of tackling online hate. This is a widely held position. For instance, Louis Reynolds of the *Institute for Strategic Dialogue*, comments, ‘Rather than solely focusing efforts to stop young people coming into contact with these views, we need to give them the critical thinking and media literacy skills to see through them’ [3]. This focus on media- and digital- literacy is understandable given the societal benefits it offers beyond just challenging online hate, as well as the minimal risk of other negative consequences from giving people more online skills and knowledge [4].

However, we draw attention to several challenges in viewing media literacy as a 'solution' to online hate. First, to our knowledge there are no quantitative longitudinal studies which measure how media literacy changes people's ability to identify and challenge the harmful effects of hateful content. This evidence base is sorely needed, especially given how much focus there is on media literacy. Second, it is likely that media literacy will work better on some people than others. For instance, it may do little to challenge deeply committed purveyors of hate but it could be highly effective at enabling some people to avoid hate-filled 'rabbit holes'. But this also needs to be explored more; literacy is only a viable solution if the problem is lack of knowledge/critical skills. This might be a somewhat naïve position to take with regards to the spread of hateful messages, especially given evidence that a substantial number in the UK have attitudinal affinity with far right ideas 5]. Third, media literacy can mean different things in different contexts and there is a need for what counts as 'provision of media literacy measures and tools' to be clearly defined.

Finally, a promising way of tackling online hate is through sophisticated safety tech, such as user-enabling (and user-controlled) information systems. In this regard, we highlight the BBC's *Own It App* as an example of 'nudge' based safety technology, aimed at minors, which has had a positive response [6]. The app is a keyboard plugin which gives users (children) nudges about their messages, using a lightweight machine learning model. For instance, if a user types out a hateful message then this will be flagged and they are warned about how it might negatively impact the recipient. The app covers a range of harmful behaviours, including aggression and bullying. It does not store any data centrally and parents are not given access. And, importantly from the vantage of free expression, the users are not stopped from doing anything; they are only given warnings. This app, although not directly related to VSPs, shows the potential of user-enabling tech to support minimization of harmful content. It would benefit VSPs if startups, academics, and other providers could be further incentivised and supported to develop safety technology which is privacy-protecting, user-enabling and free.

[1] https://about.fb.com/news/2020/04/covid-19-misinfo-update/

[2] https://www.bbc.co.uk/news/technology-53581771

	[3] https://www.isdglobal.org/defeating-hate-speech-online/ [4] https://www.ofcom.org.uk/research-and-data/media-literacy-research [5] https://www.tandfonline.com/doi/abs/10.1080/01402380902779063 and https://www.hopenothate.org.uk/wp-content/uploads/2020/02/state-of-hate-2020-final.pdf [6] https://www.bbc.com/ownit/take-control/own-it-app
Question 20: What examples are there of measures which have fallen short of expectations regarding users' protection and why? **Please provide evidence to support your answer wherever possible.**	Confidential? – Y / N
Question 21: What indicators of potential harm should Ofcom be aware of as part of its ongoing monitoring and compliance activities on VSP services? Please provide evidence to support your answer wherever possible.	Confidential? – NO. The idea that hate can inflict harm has attracted considerable academic debate, with many disagreeing about the circumstances in which hate is harmful and *how* it inflicts harm [1-3]. Many people view 'hate' as simply a matter of offence (and therefore a question of opinion and preference) rather than harm [4]. Yet there is strong empirical evidence that hate causes real harm to people, as a 2020 scoping paper from the Law Commission on online abuse indicates [5]. We have identified five aspects of harm which arise from online hate that need to be considered. 1. *Immediate harm experienced by victims* – the experience of being targeted by online hate can cause anxiety and fear. This is heightened by hate which is personally targeted at the victim, such as a direct threat to their wellbeing. 2. *Longer term harms experienced by victims* – victims can experience mental health problems due to attacks. It can also impact their life in other ways, with some reporting that they become scared of leaving their homes following online hate [10]. 3. *Offline attacks, violence and other forms of harm* – online hate can, in some circumstances, create a febrile atmosphere and lead to offline attacks and other forms of violence. It is particularly concerning when users make threats against others. However,

whilst this is a large risk and numerous anecdotal stories of online hate leading to offline attacks have been reported, there is limited causal evidence and the link between online hate and offline harm is unclear [11].

4. *Social tensions and retaliatory attacks* – online hate can stir up social tensions, and may even motivate retaliatory attacks, creating a cycle of 'cumulative extremism' [12].
5. *Access and exclusion* – people who have been targeted by online hate report feeling unwelcome and excluded from online spaces and opportunities. The creation of a hostile atmosphere through online hate can be a powerful barrier to creating fair, accessible and inclusive spaces.

The impact of online hate, and as such the level of harm that it inflicts, depends not only on what is said but also the context in which it is spoken. Two frameworks, one provided by Susan Benesch (researcher at the University of Harvard) [6] and one provided by the UN's Rabat plan [7], explicate *how* the context matters. Benesch highlights five aspects which can make language 'dangerous':

1. The speaker – powerful figures have more rhetorical power and so any hate they spread can have far greater impact.
2. The audience – some audiences are more receptive to hateful ideas and more likely to act upon them. Some vulnerable people are affected more than others by online hate, reflecting both their personal- and social- circumstances.
3. The speech act itself – what is said/shared. Some content is intrinsically more incendiary and aggressive than others.
4. Social and historical context – certain time periods have heightened tensions, such as the aftermath of a terrorist attack.
5. Mode of dissemination – how content is shared can impact its tone and reach. This echoes the idea that 'the medium is the message'.

These factors can be used to understand the likely harm that hate will inflict. It also offers a critical perspective on some of the policy choices by big platforms. For instance, Twitter has controversiallys allowed tweets to remain online which contravene its Terms of Service if there is a 'public interest' [8]. Usually, this means tweets from politicians with large followings who engage in hate speech, personal attacks and may spread misinformation. In this case, the platform is actually allowing content to stay online

which is arguably more 'dangerous' because the speakers have large receptive followings and media exposure.

Unpicking *how* hate leads to harm can also help platforms to develop better reporting processes and metrics. For example, Facebook has proposed a viewership metric for harmful content (which, somewhat confusingly, it calls 'prevalence'). This complements its existing metrics on how much online hate is *posted* by also showing how much it is *viewed* [9]. This is a welcome addition to its reporting arsenal given that hate causes harm when it reaches audiences – so knowing how much it has been viewed is crucial.

Online hate causes hate in myriad ways, and untangling when and where it harms is always difficult. We caution that there is no easy answer to the question of 'how hate inflicts harm'. Nonetheless, Ofcom could help this situation by using regulation to ensure that platforms share more information about how they make decisions about harmful content, especially *why they make decisions* – we need more information on the rationale (and the processes behind) why some content is labelled harmful. Ofcom could also motivate platforms to provide more useful metrics, such as the number of views of online hate rather than just the number of hateful posts.

[1] https://www.hup.harvard.edu/catalog.php?isbn=9780674416864&content=reviews
[2] https://philpapers.org/rec/SIMDHA
[3] https://link.springer.com/article/10.1007/s10677-019-10002-0
[4] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2703484
[5] https://s3-eu-west-2.amazonaws.com/lawcom-prod-storage-11jsxou24uy7q/uploads/2020/09/Harmful-Online-Communications-Consultation-Paper-Summary-1.pdf
[6] https://dangerousspeech.org/
[7] https://www.ohchr.org/Documents/Issues/Opinion/SeminarRabat/Rabat_threshold_test.pdf
[8] https://help.twitter.com/en/rules-and-policies/public-interest
[9] https://transparency.facebook.com/community-standards-enforcement#hate-speech
[10] https://tellmamauk.org/wp-content/uploads/2018/07/EXECUTIVE-SUMMARY.pdf
[11] https://academic.oup.com/bjc/article/60/1/93/5537169

	[12] https://www.demos.co.uk/files/Demos%20-%20Cumulative%20Radicalisation%20-%205%20Nov%202013.pdf
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? **Please provide evidence to support your answer wherever possible, including any age-related considerations.**	Confidential? – Y / N
Question 23: What challenges might VSP providers face in the practical and proportionate adoption of measures that Ofcom should be aware of? **We would be particularly interested in your reasoning of the factors relevant to the assessment of practicality and proportionality.**	Confidential? – Y/N
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?	Confidential? – NO With regards to hate, the content hosted by platforms can broadly be split into three areas. The vast majority of content they host will be entirely *non-hateful* and a small amount will be *clearly hateful* (e.g. language that is dehumanizing, contains threats or makes demonizing and derogatory statements). Both types are relatively easy to handle (leave up and ban, respectively) and are generally easier to detect and classify with computational methods. The third type of content is what falls into the 'grey area'; either it contains *subtle* forms of abuse, such as dog-whistles, or contains a negative generalisation that is unlikely to incite violence and so is not *deeply* hateful. There is far less agreement about how 'grey area' content should be defined, detected and dealt with. Cross-industry standards do not exist for this legal but harmful content, with no shared taxonomies, guidelines or standards

currently in use. There is substantial duplication, uncertainty and inconsistency across platforms, which largely create and enforce their own policies in isolation.

Moderating 'grey area' content heightens the tension in protecting free speech versus protecting users from harm. On the one hand, such content may inflict harm and could be seen as undesirable to many people, but on the other hand it is not necessarily inflicting harm to any particular user (see our answer to Question 21) and removing it could unnecessarily limit a user's freedom of expression. As a society, we are unlikely to reach consensus on this issue given that different people have different perceptions of hate, especially for more ambiguous varieties, but nonetheless more clarity and consistency from platforms should be directed by the regulator [1].

In trying to protect users from harm, platforms will sometimes remove/flag/quarantine content that should be left online. Whilst all efforts should be taken to minimise such errors, of equal importance is whether and to what extent users can challenge moderation decisions. A well-governed platform is one that explains to users why their content was taken down, allows them to easily/quickly challenge the takedown and then reinstates some content when appropriate. Reinstatements are important because a) they give users confidence in the process and b) they show the platforms really are engaging with 'grey' content and not only moderating overt hate. Reinstatements should be done in a timely manner (e.g. 24 hours) to minimize how long users' freedom of expression is curtailed.

Measuring whether platforms truly enable users to challenge content takedowns is difficult; setting a benchmark for how much content should be challenged/reinstated (e.g. a percentage threshold) could create perverse incentives, such as platforms intentionally over-penalising to ensure a proportion is challenged and reinstated. Whether or not this actually happens, ill-thought-through metrics could undermine support for the regulatory regime.

Given this, we suggest that the following metrics are useful for monitoring and evaluation:

1. The median time taken for users to lodge a challenge against content takedowns (e.g. 1 minute).
2. The mean time for the platform to make and give effect to a decision about content challenges (e.g. 24 hours).

	3. Whether content takedowns are explained to users (yes/no). 4. The % of content that is reinstated (this should be collected solely for monitoring purposes and targets should not be set). The regulator should support this work by establishing industry-wide reporting metrics. This is a rare case where this is appropriate; usually, the challenge with establishing metrics for all platforms is that they have very different designs and so one-size-fits-all metrics can lead to uninformative data. Each platform moderates different types of content and so legitimately has different reporting practices and should be evaluated differently. For instance, Twitters' moderation policy focuses primarily on users (with some moderation of posts), YouTube's on both videos and comments, and Facebook on posts (mostly, with some moderation of users). However, with regard to a user challenging moderation decisions these design constraints do not operate in the same way. All platforms can meet the same four metrics outlined above as all should have a mechanism for users to challenge moderation decisions. And, fundamentally, their different designs make little difference to the feasibility/cost of these high-level metrics. As such, whilst in general we caution against setting industry-wide metrics, for issues like this there is a clear argument in favour. Another case where Ofcom could request more information from platforms, and stipulate metrics, is with regards to the accuracy and performance of their hate detection software. For instance, Ofcom could request internal monitoring information to be made public or it could request that performance is tested against academic benchmark datasets. This would drive innovation by identifying potential weaknesses in systems. It would also help users to understand how often the software makes incorrect results, thereby evidencing its impact on freedom of expression. [1] https://dl.acm.org/doi/abs/10.1145/3295750.3298954
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).	Confidential? – Y / N

Please provide evidence or analysis to support your answer wherever possible, including consideration on how this requirement could be met in an effective and proportionate way.	
Question 26: How might Ofcom best support VSPs to continue to innovate to keep users safe?	Confidential? – NO There are two main sources of innovation within VSPs: (1) research and development undertaken by platforms, much of which is proprietary and not shared externally and (2) product development by safety tech firms and some academics. Notably, the UK has a growing safety tech industry, with companies innovating to create new technology to tackle online harms [1]. Many of these companies are startups who work directly with both small companies and the 'big' tech players. However, despite the growth in this sector, all startups face three basic challenges: 1. The biggest social media companies have large research budgets and their products/tools are often far more technically advanced than the software that academics and small companies work with. 2. Moving from a prototype (e.g. a product at TRL 2-4) to a commercial enterprise-level product is difficult. Many companies have innovative ideas but struggle to deliver scalable and cost-effective products. Delayed or poor delivery can subsequently undermine their credibility in the marketplace. 3. Tech companies have access to the originally moderated content, which tends to be the most harmful – and therefore the most valuable for training better detection, control and support systems. Platforms largely do not make such content available to third parties. If they do not make content available then third parties can only use their publicly available data, which is likely to be less harmful/hateful. This naturally limits the quality and innovativeness of the products and systems which third parties can create. Ofcom should aim to support projects and initiatives which address these challenges, and which encourage platforms to both innovative internally and to work with third parties. For instance, DCMS has recently announced a new fund to explore data sharing infrastructure and collaboration [2]. Such efforts are a promising start and could be supported

	by regulatory initiatives to enable more sharing and collaboration across the industry. One promising example of cross-industry collaboration is in the area of terrorism and extremism, where the Global Internet Forum to Counter Terrorism has attracted support from major tech firms (e.g. Facebook, Microsoft, Twitter, YouTube) to work together to tackle terrorist content [3]. Initiatives like this could be replicated in online hate, also including startup tech safety firms in the mix. Ofcom could also use the model adopted by DASA (in the Defence and Security space), in which competitions are run to kickstart and fund innovation within key strategic areas [4]. A similar model is provided by ARPA and DARPA in the US, which the UK Government has expressed interest in replicating [5]. The Turing has also responded to a consultation on this issue, and we can provide further recommendations if needed. Beyond funding, Ofcom should continue to provide 'soft' forms of support, such as hosting events and convening workshops between key stakeholders in the market. [1] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/887349/Safer_technology__safer_users-_The_UK_as_a_world-leader_in_Safety_Tech.pdf [2] https://www.gov.uk/government/news/government-publishes-new-strategy-to-kickstart-data-revolution-across-the-uk [3] https://www.gifct.org/about/ [4] https://www.gov.uk/government/collections/apply-for-funding [5] https://committees.parliament.uk/work/265/a-new-uk-research-funding-agency/publications/
Question 27: How can Ofcom best support businesses to comply with the new requirements?	Confidential? – Y / N
Question 28: Do you have any views on the set of principles set	Confidential? – NO

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?	A key challenge facing platforms will be how to arbitrate on difficult content which falls in the 'grey area' (see our answer to Question 24). Through our work we have identified several issues with online hate which platforms should establish guidelines for. Where platforms 'draw the line' will depend on their standards, norms and the user base they attract. However, we would encourage for platforms to be explicit about their position on these issues. This will provide users with far more clarity and support robust public discussions about the decisions that platforms take. It will also enable regulators to understand whether platforms are either over-penalising content or under-protecting users. Finally, more clarity about what the standards are will let the regulator (and others) understand whether platforms are reliably, fairly and accurately enforcing them. This is important given that it is unclear whether platforms always meet their own standards [1]. 1. *Truth and 'validity'* – prejudice is sometimes expressed through psuedofactual statements which derogate a group (e.g. "X% of group X are terrorists" or "X% of group Y support FGM"). Such statements are easily weaponised to attack the group, such as by drawing conclusions about their motives and how they should be tackled. This can be highly harmful when such claims are repeated frequently. The use of psuedofactual claims to express hate raises several challenges: (1) Are psuedofactual statements by themselves hateful enough to be moderated or are they just legitimate political discourse?, (2) Does the actual 'truth' of the content make any difference to whether it is hateful?, (3) if so, How do platforms evaluate 'truth' about such complex issues and (4) Does it matter whether the speaker believes (or not) that the content is truthful when they shared it? A further technical challenge is (5) fact-checking tools are often inaccurate, raising the risk of more 'noise' in how this issue is tackled. 2. *The identity of the speaker* – who speaks is a key issue in online hate (cf. the work of Benesch, discussed above). This is particularly important with the use of pejorative terms and slurs. If the term "N*gga" is used by a black person it has a fundamentally different resonance to a white person using it. To view one as equivalent to the other could lead to unfair and restrictive outcomes, such as, in this example, labelling colloquial discussions amongst Black communities as hateful.

If such content is then banned it would mean that the communities which online hate moderation is meant to protect are, in a darkly ironic twist, being harmed. This is a well-established issue in computational research [2] and platforms should be clear about how they address this issue, including what processes they have put in place.

3. *Self-hatred* – some individuals express genuine hatred against their own identity. This is different from point (2) as in that case the use of a hateful term was rendered unhateful due to the identity of the speaker. In this case, hate is still expressed but it is directed against one's own group. Platforms should be clear whether they view this as equivalent to other forms of hate. On the one hand it can be equally harmful to other people from that group who view it. But on the other, treating it as hate could constrain freedom of expression and critical debate about problems within certain communities. This issue is likely to intersect with point (1) about the supposed 'truth' of content.
4. *Humour and irony* – Many Internet subcultures are famous for various forms of offensive, trolling and tongue-in-cheek content, often referred to as 'shit posting' [3]. Jokes and ironic statements are particularly difficult to interpret as there is always a question mark over the speakers' aim. In some cases, jokes are used to lampoon and discredit prejudicial and hateful ideas – in which case there is a strong case for leaving them online. Yet in others, they are simply ways of expressing genuine hate by belittling or showing contempt for a mocked group [4]. Whether a joke is genuinely a joke depends on both its content and how it is used, and it is often surprisingly difficult to be sure of what is expressed. Platforms need to establish clear rules around what is considered legitimately humorous/ironic versus what is considered hate – and how they make such distinctions. This will largely come down to the *intent* of the speaker, which may be hard to discern and require platforms to specify further guidelines.

Ultimately, it is important that an <u>actionable</u> set of principles are adopted by platforms to govern how they tackle online hate. They need to go beyond just providing single line definitions and, instead, should offer guidelines with examples and details of what is acceptable/not. They should also specify how definitions and guidelines have been created, such as through outreach with academics, community groups and others. Ofcom has an important role to play in supporting this work and establishing frameworks

	and expectations for platforms to be more open and transparent. Providing a laundry list of issues (as we have here) may be unhelpful given that the landscape of online hate is constantly evolving – and there are important differences between platforms. What would be more helpful, and less likely to become out of date, is for the regulator to set a requirement that platforms (a) identify the most pertinent difficult issues they face and then (b) stipulate clear policies and (c) provide working examples and user-friendly explanations. [1] https://www.stopfundingfakenews.com [2] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0237861 [3] https://firstmonday.org/ojs/index.php/fm/article/download/10108/7920 [4] https://www.aclweb.org/anthology/W19-3509/