Your response

Questions for industry	Your response
Question 1: Are you providing a UK-established service that is likely to meet the AVMSD definition of a VSP? **Please provide details of the service where relevant. The establishment criteria under the AVMSD are set out in annex 5.**	Confidential? – Y / N
Question 2: Is your service able to identify users based in specific countries and do you provide customised User Interfaces (UI), User Experience (UX) functionality or interaction based on perceived age and location of users?	Confidential? – Y / N
Question 3: How does your service develop and enforce policies for what is and is not acceptable on your service? (including through Ts&Cs, community standards, and acceptable use policies) **In particular, please provide information explaining:** □ **what these policies are and whether they cover the categories of harm listed in the AVMSD (protection of minors, incitement to hatred and violence, and content constituting a criminal offence – specifically Child Sexual Exploitation and Abuse, terrorist material, racism and xenophobia);** □ **how your service assesses the risk of harm to its users;** □ **how users of the service are made aware of Ts&Cs and acceptable use policies; and** □ **how you test user awareness and engagement with Ts&Cs.**	Confidential? – Y / N
Question 4: How are your Ts&Cs (or community standards/ acceptable use policies) implemented? In particular, please provide information explaining: □ **what systems are in place to identify harmful content or content that may breach your standards and whether these operate on a proactive (e.g. active monitoring of content) or reactive (e.g. in response to reports or flags) basis;** □ **the role of human and automated processes and content moderation systems; and**	Confidential? – Y / N

□ **how you assess the effectiveness and impact of these mechanisms/ processes.**

Question 5: Does your service have advertising rules?

Confidential? – Y / N

In particular, please provide information about any advertising rules your platform has, whether they cover the areas in the AVMS Directive, and how these are enforced. See Annex 5 for a copy of the AVMSD provisions.

Question 6: How far is advertising that appears on your service under your direct control, i.e. marketed, sold or arranged by the platform?

Confidential? – Y / N

Please provide details of how advertising is marketed, sold and arranged to illustrate your answer.

Question 7: What mechanisms do you have in place to establish whether videos uploaded by users contain advertising, and how are these mechanisms designed, enforced, and assessed for effectiveness?

Confidential? – Y / N

Question 8: Does your service have any reporting or flagging mechanisms in place (human or automated)?

Confidential? – Y / N

In particular, please provide information explaining:

□ **what the mechanisms entail and how they are designed;**

□ **how users are made aware of reporting and flagging mechanisms;**

□ **how you test user awareness and engagement with these mechanisms;**

□ **how these mechanisms lead to further action, and what are the set of actions taken based on the reported harm;**

□ **how services check that any action taken is proportionate and takes into account Article 10 of the European Convention of Human Rights (freedom of expression);**

□ **how users (and content creators) are informed as to whether any action has been taken as a result of material they or others have reported or flagged;**

□ **whether there is any mechanism for users**

(including uploaders) to dispute the outcome of any decision regarding content that has been reported or flagged; and

- □ **any relevant statistics in relation to internal or external KPIs or targets for response.**

Question 9: Does your service allow users to rate different types of content on your platform?

Confidential? – Y / N

Please provide details of any rating system and what happens as a result of viewer ratings.

Question 10: Does your service use any age assurance or age verification tools or related technologies to verify the age of users?

Confidential? – Y / N

In particular, please provide information explaining:

- □ **how your age assurance policies have been developed and what age group(s) they are intended to protect;**
- □ **how these are implemented and enforced;**
- □ **how these are assessed for effectiveness or impact; and**
- □ **if the service is tailored to meet age-appropriate needs (for example, by restricting specific content to specific users), how this works.**

Question 11: Does your service have any parental control mechanisms in place?

Confidential? – Y / N

In particular, please provide information explaining:

- □ **how these tools have been developed;**
- □ **what restrictions they allow;**
- □ **how widely they are used; and**
- □ **how users of the service, and parents/ guardians if not users themselves, are made aware of and encouraged to use the parental control mechanisms that are available.**

Question 12: Does your service have a complaints mechanism in place? Please describe this, including how users of your service can access it and what types of complaint they can make.

Confidential? – Y / N

In particular, please provide information explaining:

- □ **any time limits for dealing with complaints;**

□ how complainants are informed about the outcomes of complaints;

□ any appeals processes, how they work, and whether they are independent from the complaints processes; and

□ the proportion of complaints which get disputed or appealed.

Question 13: What media literacy tools and measures are available on your service? Confidential? – Y / N

In particular, please provide any relevant information about:

□ how you raise awareness of media literacy tools and measures on your service;

□ how you assess the effectiveness of any media literacy tools and measures provided on your service; and

□ how media literacy considerations, such as your users' ability to understand and respond to the content available to them feature in your thinking about how you design and deliver your services, for example in the user interfaces, flagging content and use of nudges.

Question 14: Do you publish transparency reports with information about user safety metrics? Confidential? – Y / N

Please provide any specific evidence and examples of reports, information around the categorisation and measurements used for internal and external reporting purposes, and whether you have measures in place to report at country/ regional level and track performance over time.

Question 15: What processes and procedures do you have in place to measure the impact and effectiveness of safety tools or protection measures? Confidential? – Y / N

If not already captured elsewhere in your response, please provide information relevant to all of the measures listed above explaining:

□ how you test and review user awareness and engagement with each measure (including any analysis or research that you would be willing to share with Ofcom);

□ how often policies and protection measures are reviewed, and what triggers a review; and

□ how you test the impact of policies on users and the business more generally, such as how you balance the costs and benefits of

new tools.

Question 16: How do you assess and mitigate the risk of inadvertent removal of legal or non-harmful content?

In particular, please provide any information on:

□ **how freedom of expression is taken into account during this assessment;**

□ **how appeals are handled and what proportion are successful; and**

□ **audits of automated removal systems and, if you have them, any metrics that relate to their effectiveness.**

Confidential? – Y / N

Question 17: Have you previously implemented any measures which have fallen short of expectations and what was your response to this?

Please provide evidence to support your answer wherever possible.

Confidential? – Y / N

Question 18: How does your service develop expertise and train staff around different types of harm? (e.g. do you have any partnerships in place?)

Confidential? – Y / N

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?

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.

Your response

Confidential? – No

We have worked with various VSPs on implementing measures that meet the requirements of article 28(b)(3) of the AVMSD 2018, and in some cases exceed the minimum requirements.

Effective gatekeeping

- Bolstering registration and password systems is one of the most effective ways to assist VSPs. Large VSPs have dedicated considerable time and resources to enhance gatekeeping: For example, YouTube has implemented article 28(b)(3)(d) of the AVMSD 2018 through its trusted flagger system; article 28(b)(3)(d) by age-gating; and article 28(b)(3)(j) with warning interstitials. Youtube has also made significant moves in the demonetisation of content that may fall under the description of harmful content outlined in article 28(b)(1)(a),(b) and (c).

- With regards to smaller platforms, TAT helped Jihadology – the world's largest clearinghouse for jihadi primary source material and original analysis – restrict access to its primary sources to those organisations which have a formal affiliation with an academic or research institution (10th April 2019, Tech Against Terrorism: Press release – Launching an updated version of Jihadology to limit terrorist exploitation of the site: https://www.techagainstterrorism.org/2019/04/10/press-release-10th-april-2019-launching-an-updated-version-of-jihadology-to-limit-terrorist-exploitation-of-the-site/). We also developed warning interstitials for Jihadology to be displayed when accessing harmful content, and restricted visibility of original source URLs and imagery for non-registered users.

Terms and conditions

- We have worked with VSPs in order to encourage robust changes pursuant to Article 28(b)(3)(a) and (b) of the AVMSD 2018 in relation to terms and conditions.
    - This has largely been a product of our mentorship programme, through which we have advised platforms including Pinterest, Mailchimp, SoundCloud, TikTok, Discord, and Cloudinary on including prohibitions against the dissemination of terrorist and violent extremist content.
    - One of our mentees, TikTok, drafted its own definition of harmful content that exceeds the current legal minimum requirements and government guidance .
- We encourage the inclusion of an explicit prohibition of harms including terrorism **and** violent extremism. We are reassured that such terms are increasingly used across the tech industry.
- We note that a standard definition or policy is not desirable or practical due to the different business models and user expectations that govern different platforms.
    - We would emphasise that many UK platforms have

already employed policies to combat terrorist and violent extremist content that go far beyond any legal requirement.

  - We recommend that any standard definitions employed should acknowledge the diversity of business models and user expectations for online speech.

  - We recommend that the regulator ensures that legal speech isn’t threatened or diminished by an expansive definition of terrorist and violent extremist content.

  - We recommend that a standard definition of terrorist and violent extremist content should derive directly from the Government’s Proscribed List of Terrorist Organisations and no further.

Terrorist Content Analytics Platform (TCAP)

- Since 2019, we have been developing the Terrorist Content Analytics Platform (TCAP) in response to Article 28(b)(3)(g) and (j) of the AVMSD 2018 regarding identifying harmful content and media literacy respectively.

  - The TCAP is a secure online

platform that automates the detection and analysis of verified terrorist content on smaller internet platforms. This will represent the world's first and largest structured dataset of verified terrorist content.

- The TCAP will support smaller tech companies in improving content moderation decisions. Often the smallest platforms have limited resources to do this on their own. The platform will also facilitate secure academic research and analysis of terrorist use of the internet using the latest methodologies from advanced analytics and data science. This will help increase understanding of the threat and identify ways to improve the global response.

- Lastly, the TCAP will augment efforts to use artificial intelligence (AI) and machine learning to detect terrorist content at scale. The platform will be available for use by vetted tech companies and academics, and will include oversight mechanisms to ensure content accuracy

- In 2020, we consulted with experts from tech companies, academia, and civil society to seek further input. The outcome of this consultation was a decision to extend the remit of TCAP so that it includes content from far-right violent extremists groups. We hope to launch a beta version of the platform in the Autumn of 2020.

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

We would urge Ofcom to monitor content and accounts that praise both designated terrorist groups and violent extremist groups.

Content can be broadly grouped into visual content and textual content. Visual signifiers include logos, branding and symbols adopted by terrorist and violent extremist groups, which are used by individuals to identify sympathetic accounts or sympathetic users in chat functions. Tech Against Terrorism use resources from civil-society organisations such as Anti-Defamation League (ADL), Southern Poverty Law Center (SPLC), Center for Analysis of the Radical Right (CARR), and the directory of far right symbols created by the [redacted]

Regarding textual content, Ofcom should afford serious attention to manifestos and testimonies that relate to future or previous attacks, such as Christchurch or Halle (both of which are increasingly promoted as videos rather than texts).

In order to ensure effective monitoring of Islamist terrorist and violent extremist content, we recommend that the regulator invests in Arabic language proficiency, and engagement with research departments and civil society organisations such as the ICSR at King's College, London, Jihadiscope, and the Counterterrorism Internet Referral Unit at Europol.

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? – No

Definitional uncertainty

- Tech Against Terrorism acknowledges that there is no universal definition of terrorism. One of our observations when engaging with tech companies is that they struggle with moderating content on their sites due to this uncertainty.

- Moreover, even when content clearly depicts terrorist or violent extremist activities, it is difficult to define whether such content constitutes terrorist propaganda or newsgathering on human right abuses related to terrorism. When platforms fail to make this distinction, they are often criticised; however, as of yet there are no clear guidelines to assist platforms on how to make these decisions, particularly when their audiences are international.

- We recommend that the regulator encourage the Government to ensure designation lists are updated and robust, in order that they might form the basis of standard definitions used across the regulatory space.

- We also recommend that the regulator consults the Consolidated United Nations Security Council Sanctions list (16th August 2020, United Nations, United Nations Security Council Consolidated List: https://scsanctions.un.org/fop/fop?xml=htdocs/resources/xml/en/consolidated.xml&xslt=htdocs/resources/xsl/en/consolidated.xsl), as it provides the best framework to the international consensus on individuals and groups defined as terrorist.

Small VSPs

- We find that the VSPs that struggle to handle abuse of their platforms are overwhelmingly smaller VSPs.

  - For reference, a micro-platform consists of 1-2 staff, and a small platform can be everything upwards to a platform with 50-100 staff.

- Members of small teams are less likely to have discrete monitoring duties, and are more likely to struggle to remove content at scale in a timely fashion. This may also be true for some of the larger VSPs, where despite having a user base of millions, they have a Trust and Safety team of less than half a dozen (August 2020, TechDirt, https://www.techdirt.com/articles/20200820/08564545152/content-moderation-knowledge-sharing-shouldnt-be-backdoor-to-cross-platform-censorship.shtml).

- Video is also difficult to moderate quickly without automated systems, and small VSPs will struggle to build these on their own.

Circumventing moderation

- Terrorist and violent extremist groups are adept at circumventing moderation procedures through a range of tactics, e.g.:

  - Adopting memes or other innocuous visual media that, due to their image format and flippant context, might not get flagged despite containing terrorist imagery;

  - Employing language that is just on the right side of what is permitted by law or by terms of use;

- ◦ Migrating to new "alt-tech sites" or hijacking smaller platforms. In part, this is a consequence of mass-removal of terrorist and violent extremist groups from larger platforms.

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

We believe that the regulator must act in a manner that is consistent with the rule of law and international human rights protections. This means that the regulator should not penalise VSPs for hosting content and speech that is legal offline. Any such penalty would engender a system of censorship whereby legal speech might be removed via extra-legal means.

We note that metrics which monitor the balance of users' right to freedom of expression and duty to moderate are rarely divulged by governments. We have worked with various platforms on enhancing their transparency reporting, yet this remains a difficult task given governments do not seem to share the same commitment to transparency as they expect from industry.

The Pledge for Smaller Companies

- We instituted the Pledge for Smaller Companies in 2017 (Tech Against Terrorism – Pledge for Smaller Companies 2017: https://www.techagainstterrorism.org/m embership/pledge/), based on the GNI Principles and internationally recognised norms as articulated in the Universal Declaration of Human Rights ("UDHR"), the International Covenant on Civil and Political Rights ("ICCPR"), the International Covenant on Economic, Social and Cultural Rights ("ICESCR"), UN Security Council resolutions and documents S/RES/1624 (2005), S/RES/2129 (2013), S/RES/2322 (2016), S/RES/2354 (2017) and S/2017/375, and the UN Guiding Principles on Business and Human Rights ("UN Guiding Principles").

- With regards to freedom of expression, our Pledge cites Article 19 of the International Covenant on Civil and Political Rights (ICCPR).

- We recommend that Ofcom acknowledges the international norms that already guide best practice when determining the balance between freedom of expression and duties of moderation.

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).

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.

Confidential? – No

Redress mechanisms should be clear and readily available; however, we would emphasise that smaller VSPs will struggle to implement any such process.

We recommend that the capacity of smaller VSPs to institute redress mechanisms is taken into account, and that any penalty for the failure to introduce an effective redress mechanism should contain an exemption criterion according to the size of the VSP workforce (July 2020, Summary of Tech Against Terrorism and GIFCT webinar on accountability mechanisms for tech platforms: https://www.techagainstterrorism.org/2020/07/22/summary-of-tech-against-terrorism-and-gifct-webinar-on-accountability-mechanisms-for-tech-platforms/).

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

Confidential? – No

The tech industry has great innovative potential, however companies need the direction and focus that academia and policy professionals offer in order to use their resources efficiently.

- We recommend that Ofcom engages industry support initiatives such as Tech Against Terrorism, TCAP and the GIFCT content incident protocol in order to share collective knowledge, insights and relationships.

- We have found that companies offered access to tools and guidance produce stronger strategies against harmful content that those solely punished for failing to meet standards.

    - In particular, small VSPs will be put into a precarious position if they are subject to penalties: either they will have to disinvest in their workforce, diminishing their capacity to moderate content effectively, or they will choose to employ overly-zealous content moderation in fear of repercussions, in so doing jeopardising freedom of expression.

- A recent and instructive example of this was in 2017, when thousands of videos showing human rights abuses in Syria, as well as the channels that featured these videos, were removed by YouTube (August 2017, Witness, https://blog.witness.org/2017/08/vital-human-rights-evidence-syria-disappearing-youtube/).

Question 27: How can Ofcom best support businesses to comply with the new requirements?	Confidential? – No *Encourage media accountability* [redacted] ○ of the Christchurch terrorist incident that Twitter removed were posted by media outlets and verified accounts. ○ Despite repeated removal attempts, Facebook and YouTube encountered significant difficulties when official media outlet accounts

posted harmful content such as video of the incident and the terrorists' manifesto.

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?

Confidential? – No

Freedom of expression

- Clear commitment to the rule of law from Ofcom and VSPs should be added to the draft principles at para. 2.49, potentially as its own free-standing principle.
    - We recommend that Ofcom do not contribute to removing legal speech from the internet – this scheme should not be used as a tool that contributes to censorship creep.

Transparency

- We encourage Ofcom's commitment to regulatory transparency.

- In terms of corporate transparency, we recommend that Ofcom acknowledges the size of VSPs when they come to draft corporate transparency requirements, and approach with sufficient regard to proportionality.

Accountability

- We recommend that Ofcom should create an appeals mechanism for any decision made by Ofcom to remove legal or otherwise wrongfully-removed content.

- We recommended that Ofcom considers the Santa Clara Principles on transparency and accountability in content moderation when devising their own procedures (February 2018, *Santa Clara Principles on Transparency and Accountability in Content Moderation*, Santa Clara University High Tech Law Institute, https://santaclaraprinciples.org/).

Please complete this form in full and return to VSPRegulation@ofcom.org.uk.