Your response

Question (Volume 2)	Your response
Question 6.1: Do you have any comments on Ofcom's assessment of the causes and impacts of online harms? Do you think we have missed anything important in our analysis? Please provide evidence to support your answer.	[Is this answer confidential? No **Vol. 2 Foreign Interference Offence (FIO)**. 6P. Recommender systems are being exploited by Foreign Intelligence Services (FIS) and by sub-state proxies to disseminate content. The framing of the argument against regulation of the current recommender algorithms is potentially flawed. A valid and important framework in the context of the Online Harms Act (and other contexts – including the transnational/international dimensions) is why content is being pushed by social media platforms or placed in front of users. It is almost as important as what content is being pushed or placed in front of users because it can be stopped or minimised at source or be used for Cyber Threat Intelligence (CTI). Based on measurable studies conducted by journalists, academics, and whistleblowers it appears clear that content is being pushed by social media platforms, or placed in front of users, prompting them to engage with that content. This potentially/actively includes FIS messaging. FIS messaging seeks to negatively influence social cohesion and political systems especially on socially or politically divisive issues and/or promote views of the world at odds with those of HMG.[1] Whilst these activities are international in scope, these sets of issues affect the UK and this section of the Online Harms Act, and the Foreign Interference Offence (FIO) are valuable and needed actions.[2] Generative AI and deepfakes are part of FIS messaging and part of a wider set of difficult issues. Increasing media literacy is helpful in this and other respects but is not an answer in itself.[3] Improving media and digital literacy are two

[1] Jonathan Greenblatt's Testimony Before the House Committee on Homeland Security Examining the Domestic Terrorism Threat in the Wake of the Attack on the U.S. Capitol (April 2 2021), https://www.adl.org/resources/news/jonathan-greenblatts-testimony-house-committee-homeland-security-examining-domestic, accessed 21 February 2024.

[2] Hearing Before the U.S. Senate Committee on Homeland Security & Governmental Affairs "Social Media Platforms and the Amplification of Domestic Extremism & Other Harmful Content" Testimony of Dr. Mary Anne Franks (October 26 2021), https://www.hsgac.senate.gov/wp-content/uploads/imo/media/doc/Testimony-Franks-2021-10-28.pdf, accessed 21 February 2024.

[3] Putin's Asymmetric Assault on Democracy in Russia and Europe: Implications for U.S. National Security A Minority Staff Report, Committee on Foreign Relations United States Senate, January 10 2018, pp. 65-97, 99-139, https://www.foreign.senate.gov/imo/media/doc/FinalRR.pdf, accessed 5 November 2019.

Question (Volume 2)	Your response
	key components in resilience against FIOs/influence operations, allied to the promotion of critical thinking skills through education programs.[4] 20. Enhanced User Control. Foreign interference/political influence campaigns should not be conflated with other types of online harms like fraud, however. An actor/intent model is perhaps a better (though also potentially flawed) approach. User verification schemes can be fooled or circumvented – especially by FIS and sophisticated cybercriminals/cybercriminal gangs. The latter are increasingly organised, motivated, and professionalising their activities and business model. This includes through ransomware. In Russia especially, they have been known to work or be co-opted into working for/with FIS.[5] The way the Foreign Influence Offense (FIO) has been framed appears abstracted as though it has not yet occurred. There are detailed studies (including work conducted by the Oxford Internet Institute – OII –, Cambridge Disinformation Lab, the EUvsDisinfo initiative, and a US Senate Select Committee report in 2017) as well as UK government reports that demonstrate the material reality.[6] Intelligence-led foreign political influence campaigns are not new but what is new is coordinated state intelligence use and uses of social media for strategic messaging against Western states including the UK (especially by Russia, China, and to a lesser extent Iran). Since renewed Russian election interference operations first began in 2004 upwards of 38 nations have been targets spanning four continents: Europe, North America, Af-

[4] Some of this thinking is evidenced in a June 2020 report by the UK's House of Lords, House of Lords, Select Committee on Democracy and Digital Technologies, Report of Session 2019–21, Digital Technology and the Resurrection of Trust, 29 June 2020, p. 16-124, https://committees.parliament.uk/publications/1634/documents/17731/default/, accessed 23 August 2023.

[5] Mark Galeotti, 'Putin's Hydra: Inside Russia's Intelligence Services', https://www.ecfr.eu/page/-/ECFR_169_-_PUTINS_HYDRA_INSIDE_THE_RUSSIAN_INTELLIGENCE_SERVICES_1513.pdf. 'Threat Group Cards: A Threat Actor Encyclopedia APT group: APT 29, Cozy Bear, The Dukes', https://apt.etda.or.th/cgi-bin/showcard.cgi?g=APT%2029%2C%20Cozy%20Bear%2C%20The%20Dukes&n=1 and 'APT28', https://attack.mitre.org/groups/G0007/, 'APT29', https://attack.mitre.org/groups/G0016/. All accessed 19 July 2023.

[6] https://www.sdmlab.psychol.cam.ac.uk/research/misinformation-publications, https://www.jbs.cam.ac.uk/faculty-research/centres/cfra/conferences-events/cambridge-disinformation-summit/, https://www.oii.ox.ac.uk/, https://euvsdisinfo.eu/, accessed 21 February 2024. Putin's Asymmetric Assault on Democracy in Russia and Europe: Implications for U.S. National Security A Minority Staff Report, Committee on Foreign Relations United States Senate, January 10 2018. https://www.foreign.senate.gov/imo/media/doc/FinalRR.pdf, accessed 5 November 2019.

Question (Volume 2)	Your response
	rica, and Asia. There are also suspicions that Russia's Wagner Group has conducted operations in Latin America.[7] As well as cyberespionage/information warfare influence campaigns, money has been covertly siphoned to foreign political parties, officials, and politicians. Some of this remains undetected.[8] These too are part of active measures campaigns which increasingly leverage social media to significant effect. The 2018 Senate Select Committee on Intelligence (SSCI)-commissioned report, 'The Tactics & Tropes of the Internet Research Agency' said Russian interference was "designed to exploit societal fractures, blur the lines between reality and fiction, erode our trust in media entities and the information environment, in government, in each other, and in democracy itself".[9] This included the use of the St. Petersburg based Internet Research Agency (IRA) with funding channeled through Yevgeny Prigozhin, the head of the Wagner Group with close ties to the Kremlin. They were adept at exploiting wedge issues which mined pre-existing cracks in society as well as trying to fund and exploit agents of influence. The results have arguably and demonstrably deepened societal divisions and increased political polarization on issues such as nationalism, culture, identity politics, and immigration. The IRA, as well as the GRU and SVR used Western social media to target sections of their electorates rapidly and pro-actively and reactively at scale.[10] In the U.S. presidential election these coordinated efforts were allied to the hacking and leaking of documents following separate breaches of the Democratic Congressional Campaign Committee (DCCC) and Democratic National Committee (DNC).

[7] Renee DiResta, Kris Shaffer, Becky Ruppel, David Sullivan, Robert Matney, Ryan Fox, Jonathan Albright, Ben Johnson, 'The Tactics & Tropes of the Internet Research Agency', pp. 99-100, https://cdn2.hubspot.net/hubfs/4326998/ira-report-rebrand_FinalJ14.pdf, accessed 27 August 2019.

[8] Kylie Atwood, Michael Conte and Devan Cole, 'Russia has spent over $300 million on influencing foreign elections since 2014, US officials say' (September 13 2022), https://edition.cnn.com/2022/09/13/politics/russia-foreign-elections-influence/index.html, accessed 5 November 2023.

[9] House of Commons Foreign Affairs Committee, Guns for gold: the Wagner Network exposed, Seventh Report of Session 2022–23, 18 July 2023, p. 9. https://committees.parliament.uk/publications/41073/documents/200048/default/ , accessed 26 July 2023.

[10] Disinformation A primer in Russian active measures and influence campaigns Hearings before the Select Committee on Intelligence United States Senate 30 March 2017, pp. 2-3, https://www.intelligence.senate.gov/sites/default/files/documents/os-trid-033017.pdf, accessed 20 January 2020. See also Thomas Rid, *Active Measures: The Secret History of Disinformation and Political Warfare* (New York: Farrar, Straus and Giroux, 2020).

Question (Volume 2)	Your response
	American society was already highly polarized and politically fractured but it is argued Russian election interference in 2016 also "created a crisis of confidence in the legitimacy of outcomes and the electoral process, and exacerbated social fissures – all worthy Russian goals even if the Kremlin did not have a direct electoral impact".[11] The use of propaganda to incite or promote division and destabilization was also a feature of the Brexit referendum and other European elections or referenda including in France, Italy, Germany, Spain (especially in the Catalonian referendum of 2017), Ukraine, the Balkans and Turkey.[12] Further systematic (and ongoing) study is needed. Their ability to message is not helped by the fact that many/all of the main social media platforms have withdrawn access to their Application Programming Interfaces (APIs) and are the only ones with live data and datasets.[13] 6P10. Attributing (whether UK or allied government policy or not) is tricky less for technical or resource reasons but because of the way state-actors like Russia use proxies. This encompasses much more nuance than organised actors like the Internet Research Agency (IRA). FIS activities are multi-dimensional. Whether the full scope of these operations is captured by the act will be tested. Moreover, whether the UK's security services (were and possibly still are) fully geared to deal with current and evolving threats in this space is also a valid (but much wider) question. The framing of the offence and some parts of the Online Harms Act (because of the narrative framing) give rise to these concerns.

[11] David Gioe, 'Cyber operations and useful fools: the approach of Russian hybrid Intelligence', *Intelligence and National Security*, Vol. 33, No. 7 (December 2018), p. 956.

[12] Andrew Dawson and Martin Innes, 'The Internet Research Agency in Europe 2014-2016), Crime & Security Research Institute, Cardiff University (May 2019), https://static1.squarespace.com/static/57875c16197aea2902e3820e/t/5cd14804104c7bb3cafeaa06/1557219339758/The+Internet+Research+Agency+In+Europe+2014-2016.pdf, accessed 28 August 2019.

[13] Jessamy Perriam, Andreas Birkbak & Andy Freeman, 'Digital methods in a post-API environment', International Journal of Social Research Methodology, Vol. 23, No. 3 (2020), pp. 277-290.

Question (Volume 4)	Your response
Question 14.2: Do you have any comments on the draft guidance set out in Annex 9 regarding whether content is communicated 'publicly' or 'privately'?	*[Is this answer confidential? Yes / No (delete as appropriate)]* *General comments* We welcome the emphasis in the guidance (at para A9.19) on matters of substance, as opposed to whether the content (or parts of the service on which the content is generated, shared, or uploaded) is labelled as 'private'. The guidance also states that the fact that content has been generated, shared, or uploaded by a user that has anonymity or using a pseudonym is not expected to be relevant to the question of whether content has been communicated 'publicly' or 'privately' (para A9.19). There may be situations, however, in which conditions of anonymity are telling. For example, private Islamic State (IS) channels on Telegram are secretive and difficult to access. For some, qualities such as secrecy, anonymity and limited access are the hallmark of privacy. Yet there are important respects in which private IS Telegram channels are not analogous to, for example, a private group in which family members exchange messages and photos. In particular, private IS channels are characterised by anonymity, with pseudonyms or random strings of letters and numbers used for user IDs. Channel administrators will often not know the identities of the users in the channel. (It is this anonymity that enables some researchers and investigators to gain access). In reality, the difficulties in accessing these groups are designed not to limit the members of the group to trusted family and friends. Rather, it is to limit access to just one section of the public (pro-IS users). But, as the guidance states at para A9.23, where content "is accessible to a substantial section of the public, it should be considered as communicated 'publicly'."[14] The distinction between how content is communicated, and the nature of the content itself, is a useful one (para A9.15). As the guidance points out, it is possible for content that engages a person's Article 8 ECHR right to privacy to be communicated publicly. At the same time, it is also worth noting that there may be circumstances in which the nature of the content should inform the decision whether the content is communicated publicly or not. As an example, take an official newsletter or magazine of a terrorist organisation. Such a publication is produced with the express purpose that it

[14] The word 'substantial' is discussed further below.

Question (Volume 4)	Your response
	be widely circulated. When it is initially shared between group members for onward distribution, this communication (however secretive) should be viewed in its wider context. The public-facing nature of the content, and the desire to disseminate it to as wide an audience as possible, should inform the decision whether the initial communication was public or not. This raises wider questions regarding chain dissemination processes in which terrorist propaganda is disseminated to the public via a multi-step process, often involving multiple platforms.[15] The guidance offers some advice for such situations: • The fact that the initial communication is private does not mean that subsequent communications of the same content are also private (para A9.20). • It may be virtually impossible for services to prevent content from being shared or forwarded in certain ways (such as by taking a screenshot of content and then sending it to another user, or where a user has been given a password to access specific content and chooses to share that password with others). This does not indicate that content can be forwarded or shared with ease for the purpose of Factor (C) (para A9.39). The guidance also states that: (1) the more individuals in the UK are able to access the content, the more likely it is to be communicated publicly (para A9.23); and, (2) the converse is not necessarily true. As para A9.24 recognises, "The fact that it may be difficult for individuals to access the content (for example, because users need to take time to locate the content and it is not easily discoverable) does not mean that content should be considered as communicated 'privately'." This is especially relevant to the chain dissemination of official terrorist propaganda. When this content is initially released, the restrictions on access are designed to safeguard the early stages of the dissemination process and enable wider subsequent circulation of the materials. Here, the fact that the restriction serves the purpose of wider dissemination and making the content more, not less, publicly available should be taken into account. *Comments on Factor (A): Number of UK individuals able to access the content* The guidance states that content should be considered as communicated publicly where it is accessible to a 'substantial' section of the

[15] Hall, J. and Macdonald, S., 2023. Online Safety Bill: Distinguishing between public and private communication.

Question (Volume 4)	Your response
	public (paras A9.23, A9.29, A9.40). This raises the question how substantiality is to be assessed. Is it a quantitative assessment? Or a qualitative one? Or either/some combination of both? It should also be stated explicitly that the fact that content is accessible to a section of the public that is *less* than substantial does not mean that the content is not communicated publicly. A contrary position would be out of sync with the 'Encouragement of Terrorism' offence (Terrorism Act 2006, s. 1) – one of the priority offences listed in Schedule 5 of the Online Safety Act. For this offence, it is enough that a statement is published to *any* section of the public. It would be incoherent if a statement could be communicated publicly for the purposes of the Encouragement of Terrorism offence, yet be regarded as communicated privately for the purposes of the Online Safety Act. The guidance on Factor (A) does not address the accessing of content using a VPN. If a particular communication channel is geo-blocked in the UK, but not elsewhere, should UK users be regarded as unable to access the channel's content notwithstanding the possibility that it might be accessed using a VPN? This appears to be the assumption, but it would be useful to make this explicit. The importance of the words "by means of the service" should also be noted (Online Safety Act 2023, s. 232(2)(a)). Factor (A) does not simply require consideration of how many individuals in the UK are able to access the content. It requires consideration of how many individuals in the UK are able to access the content *by means of the service*. In the context of chain dissemination of terrorist propaganda, this distinction is significant when considering 'aggregator' platforms (i.e., ones that provide lists of URLs from which items of propaganda can be downloaded, to be publicised on 'beacon' platforms). An aggregator platform might play an important role in enabling numbers of UK individuals to access a propaganda item, but these users would not be accessing the content *by means of the aggregator platform's service*. It should be noted, therefore that additional factors beyond those listed in the statute may be considered relevant (para A9.18). An aggregator platform should be expected to consider the extent to which it facilitates access by UK individuals to content by means of *another* service. <u>*Comments on Factor (B): Access restrictions*</u> In the application of this factor, regard should be had not just to the technical features of the platform, but also these features' practical operation. For example, at the technical level Telegram private

Question (Volume 4)	Your response
	channels are designed to restrict users to those approved by the channel administrator. But in practice, joinlinks to private IS channels are often made openly available (albeit difficult to locate). This undercuts the raison d'etre of the privacy-enabling feature. It seems implicit in the guidance that the practical operation of restrictions on access should be considered, as well as the nature of the restrictions themselves. Nonetheless, an explicit statement to this effect would be worthwhile.
Question 14.3: Do you have any relevant evidence on: • The accuracy of perceptual hash matching and the costs of applying CSAM hash matching to smaller services; • The ability of services in scope of the CSAM hash matching measure to access hash databases/services, with respect to access criteria or requirements set by database and/or hash matching service providers; • The costs of applying our CSAM URL detection measure to smaller services, and the effectiveness of fuzzy	**Terrorism content URL detection** *Harms and risks* The guidance focuses on the use of URLs to share terrorist content. This is understandable and reflects the academic research in this area. While the use of URLs to circumvent content moderation and share content is important, it should be emphasised that terrorist groups and their supporters also use URLs in other ways and for other purposes. At paragraph 14.158, the guidance notes the risks of inadvertent viewing of CSAM. A similar point applies to terrorist content. Research has found that IS supporters use such tactics as hashtag hijacking and use of the @reply and @mention functions to increase the reach of their propaganda and expose unsuspecting users to it.[17] In a new VOX-Pol report, Stuart Macdonald and Sean McCafferty use data collected from four platforms over a two-month period to examine how a total of 796 items of jihadist propaganda were disseminated.[18] In respect of URLs, the study found that outlinking was the predominant method of content dissemination employed by Al-Shabaab, and that it was used regularly by Al-Qaeda. For Islamic State, outlinking was widely used to share videos, magazines, and instructional materials, but rarely used for other types of content such as bulletins, banners and photosets. The study also examined the use of inlinking. While this was the least commonly used method for content-sharing, the report expresses concern at the use of inlinks to create manually a filter bubble effect. Many of the inlinks that were collected were included in

[17] Mohammed Al Darwish (2019), 'From Telegram to Twitter: The Lifecycle of Daesh Propaganda Material'. VOX-Pol Blog, September 11. https://www.voxpol.eu/fromtelegram-to-twitter-the-lifecycle-of-daesh-propaganda-material/.

[18] Stuart Macdonald and Sean McCafferty (2024), *Online Jihadist Propaganda Dissemination Strategies*. VOX-Pol Research Report.

Question (Volume 4)	Your response
matching[16] for CSAM URL detection; • The costs of applying our articles for use in frauds (standard keyword detection) measure, including for smaller services; and • An effective application of hash matching and/or URL detection for terrorism content, including how such measures could address concerns around 'context' and freedom of expression, and any information you have on the costs and efficacy of applying hash matching and URL detection for terrorism content to a range of services.	posts beneath another item of content – so that after viewing one item users could then choose to view another, similar item on the same platform. In fact, some of these posts provided catalogues of similar content. This was the case, for example, with the nasheeds contained within the dataset. At a time when much concern is being expressed about the potential for algorithmic recommender systems to create echo chambers and take users down the 'rabbit hole', it is important that manual efforts to use inlinks to do something similar are not overlooked. There is also evidence that inlinks tend to be used more frequently to direct users to other *dissemination spaces* (such as channels or groups) on the same platform, as opposed to other items of *content*.[19] This exacerbates concerns about the use of inlinks to create a filter bubble effect – as consumers of such content are signposted to other dissemination outlets – and is a harm that is overlooked by an exclusive focus on the use of URLs to share content. *Freedom of expression* International human rights treaties stipulate that restrictions on the right to freedom of expression are permissible, provided that such restrictions are prescribed by law, pursue a legitimate objective, and meet the demands of necessity and proportionality. Legitimate objectives include the prevention of crime and the protection of national security. Accordingly, where URLs are being used deliberately to disseminate content that promotes or encourages terrorism, it is permissible to impose restrictions on the right to freedom of expression of the users posting these URLs. To minimise the impact on freedom of expression, in most instances the URL that is deactivated should link to a specific item of content. However, in some cases it may be justifiable to shut down a broader space, such as a channel or group that exists for the explicit purpose of sharing terrorist content. This mirrors the approach taken in the guidance to CSAM, at paragraph 14.163. At the same time, it is important to acknowledge that the effort to identify and disable these URLs does entail some risks to freedom of expression. There are several contributory factors: • The UK's statutory definition of terrorism has been widely criticised for being overly broad, including by the Supreme

[16] Fuzzy matching can allow a match between U2U content and a URL list, despite the text not being exactly the same.

[19] In addition to the report by Macdonald and McCafferty, see also Samantha Weirman and Audrey Alexander (2020), "Hyperlinked Sympathizers: URLs and the Islamic State", *Studies in Conflict & Terrorism* 43, no. 3: 239-257, https://doi.org/10.1080/1057610X.2018.1457204.

Question (Volume 4)	Your response
	Court.[20] The effect is to vest significant discretion in those applying the definition: in the current context, human moderators, or even automated tools. This creates a risk of inconsistent, possibly inappropriate, application of the definition, and raises the question whether the interference with the right is sufficiently clear to meet the prescribed by law threshold. • Tech companies may adopt a cautious approach to content moderation, in order to avoid accusations of failing to remove extremist or terrorist content from their platforms. This can result in over-enforcement. It has been argued that regulatory regimes that impose time limits for the removal of content may exacerbate this risk. • When applying prohibitions on terrorism-promoting content, many tech companies refer to a "greyzone".[21] This is particularly relevant to content posted by activist groups and movements. Here, the right to freedom of expression is especially important. These actors use online platforms to raise awareness of their cause, to coordinate their activities and to document human rights abuses. Yet it may be unclear whether such content falls within expansive definitions of terrorism. In some circumstances, prohibitions on terrorism-promoting content may even be used to silence activists and their supporters. The upshot is that the protection of freedom of expression requires the exercise of nuance and judgement in the application of prohibitions on terrorist content. While automated tools for the identification of online terrorist content are essential, their limitations must also be acknowledged. Machine learning algorithms have difficulty understanding context and accounting for such things as subtlety, irony, and sarcasm. This is particularly important for some types of content, e.g., memes. They also have difficulty making inferences of intention. Yet intention is central to definitions of terrorism. And there are linguistic and cultural limitations, such as assessing culturally shaped English usage in countries in the Global South. As well as overenforcement, the limitations of machine learning algorithms have resulted in documented failures to remove hate speech. This poses an additional risk to freedom of expression, as it has a chilling effect on the use of online platforms by the targeted

[20] *R v Gul* [2013] UKSC 64: see [28]-[29], [33]-[37] and [60]-[64].
[21] Isabelle van der Vegt, Paul Gill, Stuart Macdonald and Bennett Kleinberg (2019), *Shedding Light on Terrorist and Extremist Content Removal.* GRNTT. https://gnet-research.org/2020/01/07/shedding-light-on-terrorist-and-extremist-content-removal/.

Question (Volume 4)	Your response
	users and groups. It has also contributed to real-world violence in some instances, such as in Ethiopia and Romania.[22] One safeguard against overenforcement is to require tech companies to have publicly available definitions of terrorism that are properly circumscribed. For example, in 2019 Facebook narrowed its definition of terrorism so that, instead of referring simply to "violence against persons or property", it instead referred to violence against "civilians, or any other person not taking direct part in the hostilities in a situation of armed conflict".[23] The reason for this change was so that the definition could not be accused of including broader dissident groups or activist networks in conflict zones. A further safeguard is to ensure human-in-the-loop content moderation processes. However, companies employing human moderators should be mindful of two important considerations. The first is capacity, in terms of both volume of content and the necessary expertise. This encompasses linguistic and cultural understanding, as well as subject matter expertise. The second is health and wellbeing. Moderators have consistently reported suffering from significant mental health issues, with an absence of meaningful programmes to help address the consequences of regularly viewing large volumes of the most graphic and harmful content. Lack of the necessary capacity, expertise or wellbeing provision has been shown to have a detrimental impact on the quality of content moderation decisions.[24] It should be regarded as a systemic risk to the moderation of online terrorist content in compliance with users' freedom of expression.
Question 19.1: Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.	***On the amplification of illegal content*** We welcome the proposals in Volume 4 (19) to carry out on-platform tests to ensure the minimisation of the potential amplification of illegal content. Volume 4 (19) notes that if illegal content that is uploaded on a U2U service and is missed by content moderation systems, it could potentially be amplified and spread. It must be stressed, however, that research does not suggest that there is

[22] Stuart Macdonald, Ashley Mattheis and David Wells (2024), *Using Artificial Intelligence and Machine Learning to Identify Terrorist Content Online*. Tech Against Terrorism Europe. https://tate.techagainstterrorism.org/news/tcoaireport.

[23] Stuart Macdonald, Sara Giro Correia and Amy-Louise Watkin, 'Regulating terrorist content on social media: automation and the rule of law' (2019) 15 *International Journal of Law in Context* 183. https://doi.org/10.1017/s1744552319000119.

[24] Paul M. Barrett (2020), *Who Moderates the Social Media Giants? A Call to End Outsourcing.* NYU Stern Center for Business and Human Rights. https://www.stern.nyu.edu/experience-stern/faculty-research/who-moderates-social-media-giants-call-end-outsourcing.

Question (Volume 4)	Your response
	widespread proliferation of illegal content being amplified by recommender systems. In his literature review of 13 studies on the amplification of extremist content, Whittaker notes that while most studies analyse content that would be considered "legal but harmful" or "borderline" content.[25] In the rare cases that the studies do focus on illegal content, such as those by Murthy,[26] or Berger,[27] the data were collected in the mid-2010s when content moderation norms were substantially different online. Similarly, Yesilada & Lewandowsky's do not focus explicitly on the legality of material in their meta-analysis of YouTube's recommendation system.[28] However, most of the studies which they include appear to focus on content that would be considered legal under UK law. For example, health misinformation, pseudoscientific content, content which is unsafe for children, and extremist (but not necessarily illegal) content. They do include a category of "racist content", although it is again unclear whether this would be illegal. In an ongoing scoping review conducted by three of the authors of this response, which includes over 50 pieces of empirical research, there is again very little illegal content under study. ***On legal but harmful content and recommender systems*** Existing research points to a very modest amount of illegal content being amplified by recommender systems on U2U platforms. This means that the majority of potentially harmful content will not be covered by the Online Safety Act. As mentioned above, three of the authors of this paper are conducting a scoping review on the amplification of illegal, or legal but harmful content by recommendation systems. While only a small fraction of the 53 pieces of empirical research relates to content that would be considered illegal, 25 focused on mis/disinformation and 20 focused on extremist-related content (with a further four focusing on both). For misinformation, there were a number of studies that suggested conspiracy and misinformation content may be promoted by recommender systems, if users were seeking this content out and there was some evidence of misinformation filter bubbles. For example, Hussein et al focused on the promotion of conspiracy content and found evidence of a misinformation filter bubble for search

[25] Whittaker, J. (2022). Recommendation Algorithms and Extremist Content: A Review of Empirical Evidence. Global Internet Forum to Counter-Terrorism. https://doi.org/10.5210/fm.v25i3.10419.

[26] Murthy, D. (2021). Evaluating Platform Accountability: Terrorist Content on YouTube. *American Behavioral Scientist*, 65(6), 800–824. https://doi.org/10.1177/0002764221989774

[27] Buerger, C. (2021). #iamhere: Collective Counterspeech and the Quest to Improve Online Discourse. *Social Media + Society*, 7(4), 20563051211063843. https://doi.org/10.1177/20563051211063843

[28] Yesilada, M., & Lewandowsky, S. (2022). Systematic review: YouTube recommendations and problematic content. *Internet Policy Review*, *11*(1). https://doi.org/10.14763/2022.1.1652

Question (Volume 4)	Your response
	results.[29] Additionally, Matamoros-Fernandez et al found some misleading content is still recommended on YouTube.[30] There was also some evidence of extremism-related content being promoted in certain instances, and some evidence of echo chambers and filter bubbles. For example, Charles explored the amplification of white supremacist content by YouTube's recommender system and found that creators de-emphasised race in their content to market racism in an appealing way to mainstream politics.[31] Furthermore, Cockroft found that openly white nationalist figures were not as well represented in recommendations as political influencers whose content could be categorised as borderline hateful.[32] The two studies on disturbing and violent content found that this content can be promoted following non-disturbing children's videos.[33] Furthermore, the study focusing on eating disorder content found themes including the glorification of weight loss and food to achieve health and thinness within content in algorithmically promoted hashtags on TikTok.[34] Overall, the findings from the scoping review suggested that harmful content can promoted by recommender systems in certain circumstances, but limited amounts of this content is illegal. As such, unless the user is a young person, most of the potentially harmful content that is on platforms and can be promoted by recommender systems is not in scope to be addressed within the consultation. ***On other design features of recommendation systems***

[29] Hussein, E., Juneja, P., & Mitra, T. (2020). Measuring misinformation in video search platforms: An audit study on YouTube. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW1), 1-27. https://eslam-hussein.me/pdfs/papers/hussein_CSCW2020.pdf

[30] Matamoros-Fernández, A., Gray, J. E., Bartolo, L., Burgess, J., & Suzor, N. (2021). What's" Up Next"? Investigating Algorithmic Recommendations on YouTube Across Issues and Over Time. *Media and Communication*, *9*(4), 234-249. https://doi.org/10.17645/mac.v9i4.4184

[31] Charles, C. (2020). (Main)streaming Hate: Analyzing White Supremacist Content and Framing Devices on YouTube. [Doctoral dissertation]. University of Central Florida. https://stars.library.ucf.edu/cgi/viewcontent.cgi?article=1026&context=etd2020

[32] Cockcroft, D. (2020). UP NEXT: YouTube's Recommendation System and the 2019 Canadian Federal Election. [Master's thesis]. University of Alberta. https://era.library.ualberta.ca/items/0929b298-93c9-4ea6-9888-f4fc1d3db3c9/view/7038ab64-cf03-4950-aec0-d6925a2c667e/Cockcroft_Daniel_202006_MA-MLIS.pdf

[33] Balanzategui, J. (2023). 'Disturbing'children's YouTube genres and the algorithmic uncanny. *New Media & Society*, *25*(12), 3521-3542. https://doi.org/10.1177/14614448211049264; Papadamou, K., Papasavva, A., Zannettou, S., Blackburn, J., Kourtellis, N., Leontiadis, I., Stringhini, G. & Sirivianos, M. (2019). Disturbed youtube for kids: Characterizing and detecting disturbing content on YouTube. *arXiv preprint arXiv:1901.07046*. https://encase.socialcomputing.eu/wp-content/uploads/2019/01/DisturbedYouTubeforKids.pdf.

[34] Ávila, A. A. (2022). When the algorithm strikes against you: an analysis of the impact of diet culture content on TikTok on the development of eating disorders and body dissatisfaction among female undergraduates [Bachelor's dissertation]. Universitat Pompeu Fabra Barcelona. https://repositori.upf.edu/bitstream/handle/10230/54374/Ayguasanosa_2022.pdf?sequence=6&isAllowed=y

Question (Volume 4)	Your response
	One suggestion to address the potential amplification of illegal content is the use of algorithms to promote counter-speech. The best-known example of this is Moonshot's Redirect Method. The pilot of this campaign used google ad technology to redirect users searching for jihadist extremist content towards counter-narrative playlists in YouTube.[35] This approach was deemed somewhat effective as 500,070 minutes of video were watched by 320,906 individuals during the 8-week pilot.[36] This campaign has since been expanded to redirect users searching for far-right content, as well as other types of harmful content such as users searching for child sexual abuse materials being redirected towards support,[37] as well as being deployed on Facebook.[38] Another example of this use of algorithms is a Swedish Facebook group which utilised Facebook's commenting algorithm to amplify their comments whilst burying hateful comments as their counter-speech strategy.[39] Furthermore, the reach of #faces4heritage Facebook's page was found to exponentially increase when their posts were sponsored because Facebook's algorithm prioritised them over organic posts.[40] As such, there are a range of ways that algorithms can be used to promote counter-speech. However, counter-speech needs to be used with caution to avoid causing further harm or counter-productive impacts. For example, it is hard to know how effective the redirect method is as short-term reach and engagement metrics that are used do not provide a full picture of the long-term impact of these programmes (positive or negative).[41] Additionally, Schmitt et al found that counter-speech

[35] Moonshot. Redirect Method. Moonshot CVE. https://moonshotteam.com/the-redirect-method/

[36] Shain, J. (2017). *Anwar al-Awlaki: Tracking Google's Counter-Narrative Program.* Counter Extremism Project. https://www.counterextremism.com/anwar-al-awlaki-counter-narrative

[37] Helmus, T. C., & Klein, K. (2018). *Assessing outcomes of online campaigns countering violent extremism: A case study of the redirect method* (p. 19). RAND. https://apps.dtic.mil/sti/pdfs/AD1086558.pdf; Stop It Now. https://www.stopitnow.org.uk/concerned-about-your-own-thoughts-or-behaviour/concerned-about-use-of-the-internet/self-help/understanding-the-behaviour/images-are-children/?_gl=1*5eecp3*_up*MQ..*_ga*MTQ2Mzc5NjIyOC4xNjkxNDI-4ODY5*_ga_STZD47XNW7*MTY5MTQyODg2OC4xLjEuMTY5MTQyODg2OC4wLjAuMA.

[38] Moonshot (2020) Facebook Redirect Programme: Moonshot Evaluation, *Moonshot CVE*. https://moonshotteam.com/resource/facebook-redirect-programme-moonshot-evaluation/

[39] Buerger, C. (2021). #iamhere: Collective Counterspeech and the Quest to Improve Online Discourse. *Social Media + Society, 7*(4), 20563051211063843. https://doi.org/10.1177/20563051211063843

[40] De Ascaniis, S., Della Monica, C., & Cantoni, L. (2017). A Social Media Campaign to Raise Awareness About Violent Heritage Destruction. The Case of# faces4heritage. *Pori, Finland*, *2017*, 35. https://www.utupub.fi/bitstream/handle/10024/172088/HTHIC%202017.pdf?sequence=1&isAllowed=y#page=49

[41] Reed, A., & Ingram, H. (2019). A Practical guide to the first rule of CT-CVE messaging. https://www.europol.europa.eu/cms/sites/default/files/documents/reed_ingram-a_practical_guide_to_the_first_rule_of_ctcve.pdf

Question (Volume 4)	Your response
	videos associated with ExitUSA were connected with extremist videos within two clicks via YouTube's recommendation algorithm.[42] Zieringer and Rieger supported these findings.[43] More broadly, the efficacy of counter-narratives has been questioned. In a meta-review on this topic, Jones finds there to be little-to-no robust evaluation and as a result, none of the 139 campaigns being deemed to be effective.[44] Similarly, in a review of interventions conducted by Hassan et al. only three campaigns were found to show mostly positive results, and none of these actually measured whether viewing such a narrative had a positive effect on attitudes or behaviours, which limits the positive conclusions that one can draw.[45] More recently, studies by Carthy & Sarma,[46] and Braddock,[47] have used rigorous methodologies to assess the efficacy of counter-narratives and have both shown positive results. Importantly, algorithms cannot replace human involvement in the design of counter-speech campaigns, as automatically designing counter-speech can have harmful consequences. For example, Estrella Vallecillo-Rodríguez et al automatically designed counter-speech via natural language processing algorithms.[48] This was a time-efficient way of responding to hateful content, but it also resulted in grammatical errors and inconsistencies/ false information within message content. This is conducive to creating a say-do gap, which can have harmful impacts of further marginalising minority communities and exacerbating individuals' radical beliefs. Counter-speech campaigns need to be designed and disseminated appropriately to mitigate counter-productive impacts and careful ongoing

[42] Schmitt, J. B., Rieger, D., Rutkowski, O., & Ernst, J. (2018). Counter-messages as prevention or promotion of extremism?! The potential role of YouTube: Recommendation algorithms. *Journal of communication*, *68*(4), 780-808. https://doi.org/10.1093/joc/jqy029

[43] Zieringer, L., & Rieger, D. (2023). Algorithmic Recommendations' Role for the Interrelatedness of Counter-Messages and Polluted Content on YouTube–A Network Analysis. *Computational Communication Research*, *5*(1), 109. https://doi.org/10.5117/CCR2023.1.005.ZIER

[44] Jones, M. (2020). Through the Looking Glass: Assessing the Evidence Base for P/CVE Communications. *RUSI Occasional Paper*, *July*. https://doi.org/10.1093/jiplp/jpu004

[45] Hassan, G., Brouillette-Alarie, S., Ousman, S., Savard, E., & Varela, W. (2021). A Systematic Review on the Outcomes of Primary and Secondary Prevention Programs in the Field of Violent Radicalization. *Canadian Practitioners Network for the Prevention of Radicalization and Extremist Violence*.

[46] Carthy, S. L., & Sarma, K. M. (2021). Countering Terrorist Narratives: Assessing the Efficacy and Mechanisms of Change in Counter-narrative Strategies. *Terrorism and Political Violence*. https://doi.org/10.1080/09546553.2021.1962308

[47] Braddock, K. (2022). Vaccinating Against Hate: Using Attitudinal Inoculation to Confer Resistance to Persuasion by Extremist Propaganda. *Terrorism and Political Violence*, *34*(2), 240–262. https://doi.org/10.1080/09546553.2019.1693370

[48] Estrella Vallecillo-Rodríguez, M., Montejo Ráez, A., & Teresa Martín-Valdivia, M. (2023). Automatic counter-narrative generation for hate speech in Spanish. *Procesamiento del Lenguaje Natural*, *71*.

Question (Volume 4)	Your response
	monitoring and evaluation is essential to measure campaigns' impacts, and to adjust where necessary.

Question (Volume 5)	Your response
Question 26.3: What do you think of our assessment of what information is reasonably available and relevant to illegal content judgements?	*[Is this answer confidential? No* Illegal Judgements Guidance: Terrorism Ofcom recognises the significant impact illegal content judgements may have on the right to freedom of expression. It is worth noting from the outset that the risks posed to the protection of the right to freedom of expression are compounded in this context by: • The UK's statutory definition of terrorism. As stated above this has been widely criticised for being overly broad.[49] In addition to the risks already outlined under Question 14.3, the definition also does not specify express exemptions. Such as for advocacy, protest, industrial action, and dissent; or activities carried out during armed conflict as determined under international law.[50] • The terrorism-related offences set out in Schedule 5 of the Act, have also been subject to criticism for being overly broad and vague.[51] As stated above, broad definitions result in significant discretion vested in decision makers such as human moderators or automated tools. This can lead to the potential for an overly

[49] See footnote 22 above.
[50] For some discussion on express exemptions, see: Katy Vaughan, *The Interoperability of Terrorism Definitions: GIFCT Legal Frameworks Working Group* (2022): https://gifct.org/wp-content/uploads/2022/07/GIFCT-22WG-LF-TVEC-1.1.pdf.
[51] For example: Andrew Cornford (2020), 'Terrorist Precursor Offences: Evaluating the Law in Practice', *Criminal Law Review*, 663-685; J. Hodgson and V. Tadros (2009), "How to Make a Terrorist out of Nothing" 72 *Modern Law Review,* 984.

Question (Volume 5)	Your response
	cautious or inappropriate application of the guidance by services, which can have unintended consequences for the protection of human rights.[52] We welcome clarification in Volume 5 that what amounts to 'reasonable grounds to infer' in each instance will be dependent on the *nature and context* of the content in question, and that considerations will need to be given on a case-by-case basis. We welcome the acknowledgement from Ofcom that, "context is extremely important to a proper understanding of many offences and can be the difference between the reasonable grounds to infer threshold being met or not."[53] When examining impacts on the right to freedom of expression, the context of the expression can be as important as the content of the expression itself. However, the current proposed guidance appears to prioritise the extent to which the content has been communicated publicly or privately in the determination of whether it is terrorist in nature and therefore illegal content. This is just one factor to consider when making a contextual judgement as to the nature of the content concerned, and it is suggested here that to prioritise the public nature of the communication over other factors is overly simplistic and, in some circumstances, misguided. ***Information likely to be of use to a terrorist.*** The guidance states that in relation to the s.58 offence of collecting information likely to be of use to a terrorist, the state of mind element (knowledge) will be met as it is reasonable to infer that users are aware of the content they upload. This does not take into account contextual factors, which Ofcom have acknowledged are important. In relation to the defence, the guidance states that reasonable grounds to infer that the user had a reasonable excuse for posting or viewing the material *will not* arise where "content has been communicated to the general public." In determining whether the information in question is of its very nature likely to be of use to a terrorist, the courts have had to consider information and documents such as a list of fitness exercises,[54] and information on avoiding detection and concealing information from others.[55] Content, that posted publicly, could arguably

[52] Report of the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, "Ten areas of best practice in countering terrorism," A/HRC/16/51, December 22, 2010; BSR, 2021. "Human Rights Assessment: Global Internet Forum to Counter Terrorism.".
[53] Annex 10, A1.61.
[54] *R v Amjad* [2016] EWCA Crim 1618.
[55] *R v Muhammed* [2010] EWCA Crim 227.

Question (Volume 5)	Your response
	be more likely to be deemed lacking in a terrorist nature. In *R v Amjad*, central to the determination was that the material was identical to that attributed to a known terrorist. Ofcom's guidance provides examples of content authored by known terrorists or known to be distributed by terrorist networks. In this sense, it is worth considering that, it is in the context of privately communicated information that may more likely lead to the determination that the information was by its very nature of use to a terrorist. On the basis of the use of more private groups and channels by terrorist networks. This calls into question the emphasise placed on the public nature on the communication seemingly automatically leading to the inference that the reasonable excuse defence has not been met. ***Dissemination of terrorist publications*** The guidance states that when taking into consideration whether there are reasonable grounds to infer that the conduct and state of mind requirements of the offence have been met - if the terrorist publication has been uploaded to a location accessible by anyone – "it is reasonable to infer that it may be seen by somebody who would be encouraged to commit, prepare or instigate terrorism, and that most users posting such content would recognise this."[56] Again, this appears to prioritise the public nature of the communication when taking into consideration the context of the expression. The guidance states that recklessness may be inferred where the content meets the criteria of a terrorist publication,[57] as it 'should be assumed that the user posting would have recognised the risk in doing so'.[58] However, the guidance does not make clear the justification for the assumption here which appears to disregard the importance of the context of the expression. In addition, the guidance states that, it will be assumed unless the service has "clear evidence that the user did not" recognise the risk, but that services need not actively seek out such evidence before making an illegal content judgement.[59] This appears to err on the side of removal of content, which coupled with previously identified issues such as a broad definition of terrorism, widely drawn criminal offences, and the danger of an overly cautious approach taken by services presents risks for the protection of the freedom of

[56] Volume 5, 26.103.
[57] Annex 10, A2.44.
[58] Annex 10, A2.55.
[59] Annex 10, A2.55.

Question (Volume 5)	Your response
	expression online. A similar approach is taken in the guidance as to whether there are reasonable grounds to infer that the defence is applicable (that the user does not endorse the content or that the content does not express their views). The guidance appears to err on the side of removal by emphasising that services should "consider carefully" whether the attempts by the user to distance themselves from the content would be considered to be genuine. ***Context*** We recognise the difficulty in establishing in particular whether the state of mind and/or defence element of a criminal offence have been met in the context of moderating online content. However, we would emphasise the importance of context in addition to the nature of the content for the protection of freedom of expression. This can be illustrated by a decision by Meta's Oversight Board, which overturned Meta's original decision to remove an Instagram post encouraging people to discuss human rights concerns relating to the solitary confinement of a founding member of the Kurdistan Workers' Party (PKK)—a designated terrorist organization.[60] Whilst the public nature of the communication may be of relevance as part of a judgement of the context of the expression, the current guidance appears to prioritise this to too great an extent at the expense of other important contextual factors.

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

[60] "Oversight Board overturns original Facebook decision: Case 2021-006-IGUA," https://www.oversightboard.com/news/187621913321284-oversight-board-overturns-original-Meta-decision-case-2021-006-ig-ua/. Facebook had misplaced policy guidance including this exemption.