**Written evidence submitted by Carnegie UK (OSB0095)**

**Summary**

We welcome the establishment of the Joint Committee and the call for evidence to inform its scrutiny. We have submitted a detailed paper (attached), responding to each of the questions posed by the Committee's call for evidence. This summary note covers the main points, in particular the actions that we have identified for the Government to take forward in many priority areas. We look forward to discussing this further with the Committee and, as always, are happy to help both Parliamentarians and DCMS Civil Servants as they work towards the final Bill.

Introduction

Over the past three years, Carnegie UK has helped shape the debate in the UK on the reduction of online harm. We developed a proposal for a statutory duty of care to reduce Online Harms,[1] wrote our own draft Bill preceding the Government by a year[2] and provided much commentary - all of which can be found on our website.[3]

Our original proposal was, and remains, for social media companies to design and run safer systems – not for Government to regulate individual pieces of content. Our approach is to regulate the distribution system not censor individual items of content. Companies should take reasonable steps to prevent reasonably foreseeable harms that occur through the operation of their services, enforced by a regulator.

The proposal has been developed by Professor Lorna Woods (Professor of Internet Law, University of Essex), William Perrin (Carnegie UK Trustee) and Maeve Walsh (Carnegie UK Associate) and the wider Carnegie UK team. It draws on well-established legal concepts to set out a statutory duty of care backed by an independent regulator, with measuring, reporting and transparency obligations on the companies. The regime focuses on the outcome (harm reduction) making this approach future-proof and necessarily systemic.

We propose that, as in health and safety regulation, companies should run their systems in a risk-based, proportionate manner to reduce reasonably foreseeable harm. In taking this approach, the proposal moves away from a zero-sum game of takedown. Rooted in reasonableness and proportionality, systems-based regulation has the potential to allow less harmful content to remain online while mitigating its harmful impacts.

We welcome the Government's draft Online Safety Bill and are well aware of the complexity and challenges with which Ministers and officials have grappled in its development. However, we are concerned that the systemic approach governed by an independent regulator (evident in some parts of the Bill) has been eroded in favour of an emphasis on content, takedown and interventions by the Secretary of State. Ultimately, a failure to emphasise the role of systems - rather than targeting categories of content - makes the

1 https://www.carnegieuktrust.org.uk/publications/online-harm-reduction-a-statutory-duty-of-care-and-regulator/

2 https://www.carnegieuktrust.org.uk/publications/draft-online-harm-bill/

3 https://www.carnegieuktrust.org.uk/programmes/tackling-online-harm/

draft bill complex and risks rendering it ineffective at tackling the root causes of harm online[4] .

This evidence should be read alongside our recent blog post[5] in which we propose rebalancing the OSB regime to meet the norms of independence in media regulation by curtailing the Secretary of State's powers in favour of Parliament and the independent regulator. We will submit proposed amendments to these powers to the Committee when they are ready.

Our detailed evidence takes each of the Committee's questions in turn and, where appropriate, proposes actions for the Government and/or the Committee to pursue. We summarise these here.

**Objectives**

On international effectiveness and alignment with other countries' safety legislation, we recommend that the Government make a statement on how they plan to foster international co-operation on internet safety after the G7 Presidency. We also set out for the Joint Committee the key recommendations from our evidence to the Foreign Affairs Select Committee inquiry into "Tech and the Future of Foreign Policy", which are apposite to this inquiry.

On protections for children, we are concerned that the OSB dilutes the protections offered under the new video-sharing platform regime and seek that the Government explain the apparent reduction in protection for minors in respect of video content and adverts and delete clause 130.

On provisions for people who are more vulnerable to exploitation or harm online, we ask that the Government strengthen the risk assessment clauses (clauses 7 (8), (9), (10)) to take into account the risk profile produced by OFCOM and ensure an external perspective is brought to risk assessment. We recommend the Government also amend clause 7 to require publication of risk assessments.

We think that the duty of care, as described in the draft OSB, will only be partially effective and that the Bill needs to be simplified. We recommend that the Government should introduce a general duty of care under which all other duties should sit - we will provide proposed wording shortly - and simplify the other safety duties.

On the intended focus on systems and processes, again we think this is only partially delivered in the draft OSB. To achieve the benefits of a systems and processes driven approach the Government should revert to an overarching general duty of care where risk assessment focuses on the hazards caused by the operation of the platform rather than on types of content as a proxy for harm. We shall provide draft amendments to this effect in due course. The Government also needs to strengthen the language in clause 30 (1) to ensure the risk assessments function effectively.

4 See our initial analysis of the draft Bill: https://www.carnegieuktrust.org.uk/blog-posts/the-draft-online-safety-bill-carnegie-uk-trust-initial-analysis/

5 https://www.carnegieuktrust.org.uk/blog-posts/secretary-of-states-powers-and-the-draft-online-safety-bill/

On whether the draft OSB is a threat to freedom of expression, we have set out in our recent blog post why this is contingent on the removal of the Secretary of State's ability to interfere with, and direct, OFCOM on his own initiative.

**Content in scope**

In relation to these questions, we provide evidence to support the following recommendations regarding scope for the Government:

- to include online advertising in the scope of the OSB in order to tackle the impact of online fraud and scams, as called for by financial regulators, consumer groups and other Parliamentary committees.
- to provide much more clarity in relation to a series of questions relating to democratic content and content of journalistic importance.
- we do not agree with the exclusion of misinformation and disinformation and recommend that that the OSB connects to national security apparatus and regulates for national security with democratic oversight. The Counter Disinformation Cell should be put on a statutory footing with an obligation to report to Parliament and include OFCOM in its work. On societal harms more broadly, we recommend that the Secretary of State should indicate Priority Harms that address societal harms such as racism.
- To ask OFCOM to produce a paper to inform the scrutiny process with regard to the thresholds for physical or psychological harm in the regime.

**Services in Scope**

In this section, we make the following recommendations:

- OFCOM to designate the Children's Commissioner to review the evidence on ed-tech as part its Clause 61 review.
- The Secretary of State to ask OFCOM to give a "without prejudice" estimate of the threshold for Category 1 services by the end of 2021 to assist scrutiny.
- The Secretary of State to provide more detail on the justification for search engine exemptions, and the Secretary of State to ask OFCOM to give a "without prejudice" estimate of Category 2A and 2B thresholds by the end of 2021 to assist scrutiny.
- With reference to the OSB's impact on competition, that the Secretary of State provide more information on the interpretation of 'proportionate', such as reference to case law from other areas of regulation.

**Algorithms and user agency**

We urge the Government to adopt the Draft Hate Crime Code of Practice,6 drafted by Carnegie UK and other civil society organisations, which is attached to our submission and which sets out a pragmatic approach to addressing hate speech, which balances with expert human input the evident failings of pure algorithmic approaches. We also recommend that OFCOM should include adequacy of user defence tools as part of a risk assessment.

6 https://www.carnegieuktrust.org.uk/publications/draft-code-of-practice-in-respect-of-hate-crime-and-wider-legal-harms-covering-paper-june-2021/

**The role of OFCOM**

We welcome the decision to appoint OFCOM to the role of online safety regulator but call for some specific amendments to the Bill to enable it to carry out its role effectively. Much of the rationale for these relates to the concerns around the powers the OSB gives to the Secretary of State and the impact on OFCOM's independence:

- Amend clause 49(4): replace "OFCOM may only describe information of" with "OFCOM may describe such information as it sees fit to execute its duties, including"
- Remove limitation of transparency reporting in cl 49(3) (and delete 'relevant' in cl 49(1))
- In each of Clauses 9, 10 & 11 'Safety Duties' insert new final subclause: "A duty to provide the risk assessment to OFCOM when it has been carried out as described in Clause 8."
- The Joint Committee should take evidence from the National Police Chiefs Council lead on hate speech/digital issues.
- The OSB must achieve a better balance between Parliament and the Secretary of State to preserve OFCOM's independence (as described in Carnegie blog post) and amend the Bill accordingly; this would include deleting Clauses 33 and 113.
- Amend clause 109 to ensure that the strategic priorities remain high-level objectives, not control over day-to-day implementation of the regime. The Government should also explain or modify Clause 112.

On media literacy, we recommend that OFCOM should include media literacy issues in its clause 61 survey and require media literacy to be built into risk assessments as a mitigation measure. OFCOM should also use its analytical ability to estimate how many people require media literacy training and the gap between that number and current capacity. The Essential Digital Skills framework should also be reviewed to see if it is sufficient for avoiding harms.

**Carnegie UK**

**September 2021**

**Contact: maeve.walsh@Carnegieuk.org**

**Carnegie UK Trust: Response to OSB Scrutiny Committee call for evidence**

**Full response to the Committee's questions**

We address the Committee's questions below and would be happy to provide further information, in writing or in person, as the inquiry progresses.

Objectives

**Will the proposed legislation effectively deliver the policy aim of making the UK the safest place to be online?**

Many democracies are now considering how to regulate technology companies, specifically social networks[7] for their impact on society. The UK approach based on risk assessment and due diligence - which is similar to the approach in the UN Guiding Principles on Business and Human Rights[8] and OECD Guidance for Responsible Business Conduct[9] - could serve as a model for adoption by the international community.

In focussing on tools and distribution rather than content, the systems-based approach avoids some of the difficult questions about agreeing acceptable content standards across different countries; it also mitigates the impact on freedom of expression, as the Special Rapporteur for Freedom of Expression recognised in his 2019 report on hate speech[10.] The UK Presidency of the G7 achieved an extremely encouraging text on internet safety earlier this year,[11] which it is following through.

The OSB will be more effective if other countries are aligned with it and its underlying approach. We make comments on how the UK could engage internationally in response to the question on international comparator regimes below. The Government should set out to Parliament how it intends to seek international co-ordination on emerging regulatory regimes after the end of the UK G7 Presidency.[12]

*Action: Government to make statement on how they plan to foster international co-operation on internet safety after the G7 Presidency*.

---

7 OECD has produced a helpful list of the top 50 global online content sharing services https://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=DSTI/CDEP(2019)15/FINAL&docLanguage=En

8 https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf

9 https://www.oecd.org/investment/due-diligence-guidance-for-responsible-business-conduct.htm

10 A/74/486 Report to 74th Session of the General Assembly - see para 51. https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/ReportOnlineHateSpeech.aspx

11 https://www.gov.uk/government/publications/g7-digital-and-technology-ministerial-declaration

12 See our submission to the Foreign Affairs Committee's "Tech and Foreign Policy" inquiry: https://committees.parliament.uk/writtenevidence/35708/html/

**Will the proposed legislation help to deliver the policy aim of using digital technologies and services to support the UK's economic growth? Will it support a more inclusive, competitive, and innovative future digital economy?**

Yes. The market cannot allocate resources efficiently if the costs arising from very large companies' services are not incurred by them. The costs of malign aspects of social networks fall on society and the victims. For economic efficiency, the company and its shareholders should bear those costs. The OSB will ensure some external costs return to the company through improving its systems and processes on the "polluter pays" basis.[13] However, if the OSB slips into reliance on content rules and criminal offences, the burden will continue to fall upon victims (in initiating and pursuing claims) and the state (through costs in policing and court time).

Social inclusivity and individual freedoms will be harmed if people are deterred from using the services by huge quantities of harmful speech[14]. Effective regulation will empower people currently harmed by the operation of the platforms, particularly through hate speech and harassment to make fuller use of their services. Although the systems-based approach is not content regulation, the experience from broadcast, film and advertising regulation demonstrates that a skilled regulator can work to assess harm in context, regulate it and balance this with maintaining free speech.

Proportionality in regulation allows for innovation and market entry by SMEs. Moreover, the framework approach, rather than baking everything inflexibly into legislation, permits innovation and experimentation in different ways of delivering services, supporting the future-proofing of the regime.

The systemic approach could result in innovation in safety tech to improve company systems. Better safety tech might allow users to choose their own level of risk and have more control over their respective individual online experience.

A well-run, broadly based regulatory regime in the UK could in future lead to social media companies choosing to base operations here as a place from which to comply with other regimes - much as was the case with satellite broadcasting. We sketched out the advantages of a distinct British model of social media regulation in a letter to the Rt Hon Nicky Morgan MP when she was Secretary of State.[15]

13 OECD 1972

14 See recent polling from Compassion in Politics reported here: https://inews.co.uk/news/online-safety-freedom-from-abuse-more-important-than-freedom-of-speech-on-the-internet-poll-finds-1193880

15 Letter from Perrin, Woods, Walsh (December 2019): https://www.carnegieuktrust.org.uk/news-stories/letter-to-dcms-secretary-of-state-introducing-a-draft-online-harm-reduction-bill/

**Are children effectively protected from harmful activity and content under the measures proposed in the draft Bill?**

The measures to protect children appear strong, but we would defer to the views of child protection specialists - such as the NSPCC, 5 Rights and CHIS - as to whether they go far enough.

We draw the Committee's attention to a significant issue with video. The new OSB regime should not be weaker than the video-sharing platform regime[16] it replaces. That regime (not yet being enforced but which applies to SnapChat, OnlyFans and similar services[17]) requires companies to protect children from videos and adverts that might:

> *"impair the physical, mental or moral development of persons under the age of 18"*.[18]

The Government has not discussed how the new OSB regime might offer similar protection. We note both the general exclusion of advertising from the draft Bill; and a seemingly higher threshold for regulatory intervention for video directed at children in the OSB than is currently found in the Communications Act. We therefore propose the deletion of clause 130.

*Action: Government to explain the OSB's apparent reduction in protection for minors in respect of video content and adverts; and delete clause 130.*

**Does the draft Bill make adequate provisions for people who are more likely to experience harm online or who may be more vulnerable to exploitation?**

No. The measures to protect adults are inadequate (see answer below) and do not take full advantage of the systemic approach. In this regard, we question whether the risk assessment duty (in all three instances, but specifically the adults' risk assessment) have enough safeguards to ensure the quality of the risk assessment, to avoid for example, wilful blindness[19] as to the risks (we give examples of this sort of behaviour below).

Polluting companies will sometimes continue their pollution in the shareholder interest on a commercial calculus: it costs more to clean it up than it does to, say, fight it in the courts or buy counter-balancing public relations. Without regulation, internal risk assessments would then underplay the probability of harm, lack rigour or be quashed at a senior level. The tobacco industry is a distressing example[20]. The Committee should consider, however, if the mission-driven, "we're changing the world" nature of modern technology companies can

16 See Video-sharing platform (VSP) regulation, OFCOM 24 June 2021

https://www.ofcom.org.uk/tv-radio-and-on-demand/information-for-industry/vsp-regulation

17 A list of services regulated by OFCOM can be found here: https://www.ofcom.org.uk/tv-radio-and-on-demand/information-for-industry/vsp-regulation/notified-video-sharing-platforms

18 OFCOM https://www.ofcom.org.uk/__data/assets/pdf_file/0021/205167/regulating-vsp-guide.pdf

19 See for example recent evidence from Wall Street Journal that Facebook ignored signs its algorithm was "rewarding outrage": https://www.wsj.com/articles/facebook-algorithm-change-zuckerberg-11631654215

20 See WHO https://www.who.int/tobacco/media/en/TobaccoExplained.pdf

add an extra hazard to genuine internal risk assessment and mitigation through dismissal of risk in pursuit of imagined higher objectives. The Wall Street Journal[21] coverage of Instagram's internal research suggests a startling disregard for internally divulged risks.

> *'"Teens blame Instagram for increases in the rate of anxiety and depression," said another slide. "This reaction was unprompted and consistent across all groups."*
>
> *Among teens who reported suicidal thoughts, 13% of British users and 6% of American users traced the desire to kill themselves to Instagram, one presentation showed."*

In a regulated system, the formal risk assessment should be judged against external standards, not just the companies' internal judgements. The draft Bill needs to be tighter in this regard. One solution could be to amend the obligation to take into account the risk profile produced by OFCOM to read:

> *"... an assessment to identify, assess and understand such of the following as appear to be appropriate, taking due account of the risk profile ...." (cl 7(10))*

A similar amendment could be made in respect of the other two risk assessment duties (cl 7(8) and cl 7(9)).

Good corporate governance requires the publication of a range of compliance documents - most recently COVID risk assessments, but also Modern Slavery Statements - which are akin to a risk assessment of that hazard. If the service provider does not share its risk assessments with user/customers, then people will not be able to take their own mitigating steps to avoid harm. It also means that competition on safety between providers is precluded. In a wide range of consumer goods and services, providing customers with detailed information about risks is a fundamental part of doing business.

The impact of regulation in the United Kingdom will be increased if other national regulators, Parliaments and courts are able to act upon its work. The biggest social media companies also claim that they wish to see global regulatory norms. Given the reasonably uniform nature of the largest social media platforms across their markets, the risk assessment carried out in the UK will be germane in other jurisdictions.

Risk assessments of Category 1 companies should be published on the company website after having agreed with the regulator which elements are genuinely commercially confidential (it is hard though to explain why risks to the public might be commercially confidential). One way of doing this would be to amend Clause 7 to add a new sub clause:

> *"(12) Risk assessments produced under Section 7 by Category One companies should be published within six months of being made after agreeing with OFCOM the redaction of commercially confidential material."*

*Action: Government to strengthen the risk assessment clauses (clauses 7 (8), (9), (10)) to take into account the risk profile produced by OFCOM and ensure an external perspective is*

---

21 Georgia Wells, Jeff Horwitz, Deepa Seetharaman WSJ 14 September 2021 https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739?st=x2w9tezhq3gn85p&reflink=share_mobilewebshare

*brought to risk assessment. Government to amend clause 7 to require publication of risk assessments for Category 1 companies.*

**Is the "duty of care" approach in the draft Bill effective?**

It is only partially effective.

The draft Bill is hard for a lay reader to understand. This will hinder scrutiny and increase the regulatory burden. The Government should structurally simplify the Bill's three safety duties, three counterbalancing factors and its byzantine commencement process. Specifically, there should be a general safety duty to orientate and to give coherence to the regime.

We are preparing detailed amendments to the draft OSB to insert a general safety duty. This is a complex task due to the Bill's complexity; we shall submit a schedule of amendments to the Committee when ready.

Harms to adults are not well covered (clause 11). First, the obligation applies only to a limited number of services (yet to be defined, but expressly excluding search). Further, the Bill does not spell out how, for example, huge volumes of misogyny, racism antisemitism, etc – that are not criminal but are oppressive and harmful – will be addressed. We think the Government might use the "priority content" to address these but the Secretary of State has not made an indicative announcement as to what the priority harms might be. We suggest elsewhere reform to the "Priority Harms" process.[22]

Clause 11 states that services have a "duty to specify in the terms of service" how "priority content" and "other content that is harmful to adults" should be "dealt with by the services." We agree that platforms should have some flexibility in how they choose to address risks of harm arising from content or behaviour that is not criminal. Each platform should have to consider its own characteristics (business model, branding, functionalities, etc) and those of its users[23] and take appropriate action accordingly. This seems to be the policy intent. However, we have two principal concerns with the approach in the Bill.

First, "dealt with" is a phrase that has no qualitative meaning: it does not state whether it has to be done positively, negatively or by deciding not to do anything about the problem. (There is precedent for challenge of this type of language: the Irish Data Protection Commissioner is claiming that it is "handling" cases by not taking a decision on them.[24]) Comparison of clause 11 to the children's safety duty obligation to "mitigate and effectively manage" risks (cl 10(2)) demonstrates that the Government has made a deliberate choice for this to be weak, although as evidence to the Scrutiny Committee has already demonstrated, it is a highly problematic area about which there is much concern.

22 https://www.carnegieuktrust.org.uk/blog-posts/secretary-of-states-powers-and-the-draft-online-safety-bill/

23 This is a point we made in our original report: "Online Harm Reduction: a statutory duty of care and a regulator" (April 2019; p42) https://d1ssu070pg2v9i.cloudfront.net/pex/pex_carnegie2021/2019/04/06084627/Online-harm-reduction-a-statutory-duty-of-care-and-regulator.pdf FULL REF REQD

24"Irish DPC openly acknowledges: It does not decide about GDPR complaints. At least 99.93% see no decision" https://noyb.eu/en/irish-dpc-handles-9993-gdpr-complaints-without-decision

Secondly, it is important to remember that safety duties should not be just about moderation and take down. For example, a platform that wanted to adopt an "anything goes" approach, might want to ensure effective warnings at point of entry or provide their users with tools to self-curate as they adjust to risks within that online environment (some subreddits work like this). A service provider could review its design choices to understand the extent to which those choices contribute to the problem, and make changes accordingly (eg increasing friction to reduce flame wars). The extent to which the draft Bill envisages a systems-based approach is unclear. The provisions outlining the effect of the codes (cl 37) and the online safety objectives in clause 30 are based on a systems approach. The codes that OFCOM is to produce under clause 29 need only be "compatible with" the online safety objectives. The codes need not take them into account, nor need they further those objectives. In any event, it is unclear the extent to which the codes will have much impact given the weakness of the obligation on platforms in clause 11.

Clause 11 (harms to adults) relies upon platforms' enforcement of their own terms of service (as against users). In so doing, it loses close connection with the characteristics and functionalities of the platform design and operation; and their impact on content creation (e.g. through financial or other incentives), information flows (e.g. recommender systems) and user empowerment (e.g. through usable curation tools) that flows from a systemic approach. The result is that Clause 11 is too downstream an approach. By contrast, the illegal content and child safety duties emphasise the importance of these "characteristics" upstream of terms and conditions.

For a duty of care to work, there needs to be a clear understanding of the risk assessments to inform customer behaviour to make an informed choice and for the regulator to ensure reasonable steps are being taken. All the risk assessments should at an absolute minimum be shared with the regulator by default (see below) and in many cases (such as Category 1 services) be published as we note in the answer to the previous question. The obligation to notify OFCOM of the presence of non-designated content that is harmful to children (cl 7(4)) and content other than priority content that is harmful to adults (cl 7(7)) is insufficient in this regard.

We return to the duty of care in the answers to other questions. We are preparing detailed amendments to the draft OSB to insert a general safety duty. This is a complex task due to the Bill's complexity; we shall submit a schedule of amendments to the Committee when ready.

*Action: the Government should introduce a general duty of care under which all other duties should sit - we will provide proposed wording shortly - and simplify the other safety duties*.

**Does the Bill deliver the intention to focus on systems and processes rather than content, and is this an effective approach for moderating content? What role do you see for e.g. safety by design, algorithmic recommendations, minimum standards, default settings?**

Carnegie UK is an advocate of what it has termed a systems-based approach: an approach, first described by Woods and Perrin in part in 2016 working with Anna Turley MP[25] and then

set out as a full regulatory regime, firstly in a series of blog posts in 2018 then in our full 2019 reference paper.[26]

The systemic approach is valuable because social media platforms constitute artificial environments, created by someone. They have created systems which are harmful. The platforms affect the things users can do online, and also - as behavioural psychology research suggests - nudge them into behaving in certain ways. This created communicative environment is, to a large extent, the result of cumulative design choices: choices which can be pro-social or anti-social. Design choices can include algorithmic recommendations and default settings. To date, it seems that design choices on most social media and search services have been driven by the shareholder interest, irrespective of the potential consequences for users and for society. This may have been so in the early days -

> "*God only knows what it's doing to our children's brains. The thought process that went into building these applications, Facebook being the first of them, ... was all about: How do we consume as much of your time and conscious attention as possible?*" (Sean Parker, co-founder of Facebook, 2017)[27] -

and apparently continues now with Facebook's chaotic XCheck tool for 5 million VIPs that allowed them to "*violate our standards*" designed to prevent harm to others "*without any consequences*" to prevent '"*PR fires*"— negative media attention that comes from botched enforcement actions taken against VIPs.

> "'*After a woman accused Neymar of rape in 2019, he posted Facebook and Instagram videos defending himself - and showing viewers his WhatsApp correspondence with his accuser, which included her name and nude photos of her. He accused the woman of extorting him....Facebook's standard procedure for handling the posting of "non consensual intimate imagery" is simple: Delete it. But Neymar was protected by XCheck.*"[28]

This is a business choice taken without regard to the impact on victims in the individual cases, nor more broadly on the communications environment of the platform - encouraging users to disregard the rights of others as well as the platform's community standards by demonstrating people getting away with it.

Our approach often breaks down the companies' processes into a number of stages: access to the platform (e.g. privacy settings; the lack of friction in setting up accounts/replacement and coordinated accounts), creation of content (e.g. emojis, deepfake tools[29]); navigation

---

25 Malicious Communications (Social Media) Bill: https://bills.parliament.uk/bills/1877

26 https://www.carnegieuktrust.org.uk/publications/online-harm-reduction-a-statutory-duty-of-care-and-regulator/

27 Interview, Axios November 2017 https://www.axios.com/sean-parker-unloads-on--facebookgod-only-knows-what-its-doing-to-our-childrens-brains-1513306792-f855e7b4-4e99-4d60-8d51-2775559c2671.html

28 WSJ September 14, 2021, 'Facebook Documents Reveal Secret Elite Exempt From Its Rules.' https://www.wsj.com/articles/facebook-files-xcheck-zuckerberg-elite-rules-11631541353?mod=hp_lead_pos7

29 For example, the DeepSubeke Nudify App: https://www.bbc.co.uk/news/technology-57996910; see also K. Hao "A horrifying new AI app swaps women into porn videos with a click" *MIT Technology Review*, 13

and discovery (e.g. auto-completes and recommender algorithms); complaints and moderation; user self-defence (e.g. ability to change setting or select 'safe' or 'risky' experiences). Safety by design, rooted in a democratic regime should be adopted all the way through this communication process.

This systemic approach, which targets the distribution platforms and their operators, is different from simple content rules aimed at users. It opens up the debate beyond that of simply what content to take down. It can utilise other interventions that are less speech-intrusive in the first place, as recognised by the UN Special Rapporteur on Freedom of Expression in the context of hate speech[30]) who set out a useful list of options open to companies:

> *"..can delete content, restrict its virality, label its origin, suspend the relevant user, suspend the organization sponsoring the content, develop ratings to highlight a person's use of prohibited content, temporarily restrict content while a team is conducting a review, preclude users from monetizing their content, create friction in the sharing of content, affix warnings and labels to content, provide individuals with greater capacity to block other users, minimize the amplification of the content, interfere with bots and coordinated online mob behaviour, adopt geolocated restrictions and even promote counter-messaging. Not all of these tools are appropriate in every circumstance, and they may require limitations themselves, but they show the range of options short of deletion that may be available to companies in given situations."*

So rather than seeing the rights of speaker and victim as a zero-sum game, in a systemic approach other interventions may allow both to co-exist. Especially when talking about content that is harmful to adults, there is a crucial difference between the scope of the regime (should a platform be required to consider the risks) and the intensity of action required by the platform (including action other than take down). The systemic approach is much broader and, as we noted above, operates at a deeper level than just content moderation, allows greater flexibility in responses and has the potential to be effective because it looks at factors relevant to the creation of problem content and behaviour.

The draft Bill captures this in part by its focus on what it calls the "characteristics" in clause 61 and as a subset of the characteristics, the 'functionality' of services (defined in cl 135). "Characteristics" are a factor that OFCOM has to take into account in its risk assessment guidance (see cl 61). The Online Safety Objectives (cl 30), which OFCOM must be consistent with when developing codes, also refer to functionalities, as well as the algorithms used by the service amongst other factors. Clauses 61 and 30 potentially indirectly influence the risk assessment and safety duties. The wording used to describe the strength of this influence is weak, however. Clause 30(1) specifies that the steps proposed by the codes are "compatible" with the objectives, merely ensuring that there is not a conflict. We have noted the weakness in the language of the risk assessment duties, above. Further, the Bill

---

September 2021: https://www.technologyreview.com/2021/09/13/1035449/ai-deepfake-app-face-swaps-women-into-porn/

30 A/74/486 Report to 74th Session of the General Assembly - see para 51. https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/ReportOnlineHateSpeech.aspx

mixes into this a considerable amount of content-specific interventions. This mix has the following consequences:

- a complex structure, which works backwards from proxies for harm (specific categories of content) rather than forwards from the hazards created or exacerbated by the platform design and business model; and
- a focus on *ex-post* content-specific interventions that tend towards a binary choice between leaving content alone or taking it down, and which does not take advantage of the range of interventions available with a full systemic approach.

*Action: To achieve the benefits of a systems and processes driven approach the Government should revert to an overarching general duty of care where risk assessment focuses on the hazards caused by the operation of the platform rather than on types of content as a proxy for harm. We shall provide draft amendments to this effect in due course. The Government needs to strengthen the language in clause 30 (1) to ensure the risk assessments function effectively.*

**How does the draft Bill differ to online safety legislation in other countries (e.g. Australia, Canada, Germany, Ireland, and the EU Digital Services Act) and what lessons can be learnt?**

Carnegie UK has worked with a wide range of international actors on online safety - here we draw out some lessons learned from that for the United Kingdom in how it approaches international issues in regulation. This is based upon our evidence[31] to the ongoing Foreign Affairs Select Committee inquiry into "Tech and the future of UK foreign policy". The UK should:

- Demonstrate that democracies have a strong role in governing the internet instead of leaving it to global companies and unelected technologists. Baroness Kidron demonstrated the potential through 5 Rights remarkable work at the United Nations with the Convention on Rights of the Child “General Comment 25 on children's rights in relation to the digital environment”[32].
- HMG should export the UK Online Safety approach, including the statutory duty of care which bears great similarity to the due diligence obligation in the Digital Services Act, and which is being considered as a model in other Commonwealth jurisdictions.[33]
- Work towards new, strong multilateral processes for competent democratic governments to work together on technology governance embedding human rights principles, securing democratic debate and correcting market failures. The first step

31 Written evidence from Carnegie UK: https://committees.parliament.uk/writtenevidence/35708/html/

32 https://www.ohchr.org/EN/HRBodies/CRC/Pages/GCChildrensRightsRelationDigitalEnvironment.aspx

33 See for example, report from the Parliament of Victoria's Electoral Committee inquiry into the impact of social media on elections: https://www.parliament.vic.gov.au/images/stories/committees/emc/Social_Media_Inquiry/EMC_Final_Report.pdf

was the G7 tech ministers' declaration[34], secured by Oliver Dowden; and the next the recent G7 Interior and Security Ministers declaration[35]. This might require a new treaty. Damian Collins has called for a "Bretton Woods" for technology; the spirit of this is quite correct, but this should not be a UN process at the outset.

- Deploy democratic technology governance as a bulwark against autocratic technology governance – such as the China's World Internet Conference - and defend democracy itself from strategic online disinformation campaigns by hostile state actors, their proxies and fellow travellers that threaten national security. This would include developing a system for assessments of disinformation campaigns by foreign actors that threaten national security to be shared for action between the intelligence services, companies regulated under the Online Safety regime and the regulator. The OSB is practically silent on this, as we note below.
- Embrace governments that do not have the technical capacity to make their own rules in multilateral processes - similar to observer status at Basel and through systems like a reinforced Commonwealth Cyber Declaration[36] and Rule of Law programmes.
- Improve the byzantine, even chaotic UN process (WSIS,[37] etc) by external leadership that demonstrates how to do it better.
- Identify a structure within the Foreign, Commonwealth and Development Office to help DCMS manage a sustained drive of technology diplomacy over the next five years; this should include identifying who, at Ambassadorial (SMS4 level or equivalent), is responsible for the landscape of tech regulation and what resources they require.

*Action: Government to consider these approaches in forthcoming international deliberations and strategy development.*

**Does the proposed legislation represent a threat to freedom of expression, or are the protections for freedom of expression provided in the draft Bill sufficient?**

The Bill does not provide a threat to freedom of expression as long as the Secretary of State's ability to interfere with and direct OFCOM on their own initiative is removed. We demonstrated how to do this in a recent blog (attached in the annex) and will provide detailed amendments to give this effect in due course.[38]

34 https://www.gov.uk/government/publications/g7-digital-and-technology-ministerial-declaration

35 https://www.gov.uk/government/publications/g7-interior-and-security-ministers-meeting-september-2021/annex-1-statement-on-preventing-and-countering-violent-extremism-and-terrorism-online-accessible-version

36 https://thecommonwealth.org/commonwealth-cyber-declaration

37 https://www.itu.int/net/wsis/

38 https://www.carnegieuktrust.org.uk/blog-posts/secretary-of-states-powers-and-the-draft-online-safety-bill/

In summary, we suggest that a better balance can be struck between Parliament and the executive in setting priorities that maintain OFCOM's independence. We suggest examining the issue in two parts: regime start up; and response to issues during operation. The draft Bill should be amended so that:

- the Secretary of State specifies (with supporting research) the initial outcomes they seek to address and "priority content" on the face of the Bill, which Parliament can hold to account. This sets priorities during the regime start-up phase.
- during regime operation, changes to priority content should originate from OFCOM's research, not from the Secretary of State, and be rigorously evidence-based. OFCOM should form the need for new priority content from its research and its risk assessment processes (cl 61), then consult Parliament, the Secretary of State and others. OFCOM should have regard to the consultation and present a report to the Secretary of State from which they should make a Statutory Instrument (SI) (by the positive procedure) to put the new priority content into effect.

The draft Online Safety Bill envisages a continuing control in the hands of the Executive beyond high-level strategic direction. Clauses 33 and 113 affect OFCOM's role to implement policy; this should be an area in which there is no Government interference. This boundary is important in regulation of many economic sectors; it still more important where the regulation has the potential to impact on freedom of expression (even though the regime does not regulate content directly). Yet both clauses cross the boundary emphatically. Moreover, there is no attempt to provide for scrutiny or control of these powers by Parliament. The Secretary of State's power to direct OFCOM to make amendments to the code to reflect Government policy (cl 33) and to give guidance as to the exercise of functions and powers are simply egregious and should be deleted.

On freedom of expression more broadly, the draft Bill takes the British approach to balancing rights, rather than a North American approach that speech is a predominant right, favoured by many American platforms and their enthusiastic advocates. The Human Rights Act, by implementing the European Convention on Human Rights acknowledges the legitimacy of regulating mass media in the public interest[39].

The targets for regulation in mass media regimes (TV, radio, film) are also systems to distribute content to very large audiences to earn revenues. These regimes impose greater protections on content that a person might consume in their own home, and this consideration applies to social media too. They recognise that, for potentially mixed audiences, some limits on what can be said is acceptable in the public interest; and, moreover, that some content should be kept away from certain groups. The UN Convention on Rights of a Child "General Comment on the rights of the child in a digital environment" acknowledges that children can/should be protected from some types of content[40]. While these regimes are based on content regulation they still are, in principle, justifiable in the public interest. A regime such as the OSB, based on systems regulation which has a less

39Human Rights Act 1998; https://www.legislation.gov.uk/ukpga/1998/42/contents; European Convention on Human Rights (see Article 10(1)): https://www.echr.coe.int/documents/convention_eng.pdf

40 https://www.ohchr.org/en/professionalinterest/pages/crc.aspx

direct impact on expression than traditional media regulation, should be justifiable in human rights terms even more so.

In assessing a regulatory regime, it is important to check how the regime balances the different rights of the various users. It is important to remember that freedom of expression is a right that should be enjoyed equally by *all*. It is clear, however, that some groups are affected by online hate speech and abuse more than others. Consideration should be given to those whose voices have been silenced by intimidation or abuse or who have self-censored through fear of the same after seeing others like them being abused. This can be seen *in extremi*s in the Committee on Standards in Public life report on intimidation in public life[41] through its impact on politicians (and the recent Report on Women's Participation in Northern Ireland[42]) but occurs in society more broadly.[43] The OSB regime should increase protection for such people and increase their ability to express themselves. The risk assessment provisions (cl 61and cl7(8)-(10)) and the codes of practice - specifically the online safety objectives (cl 30) - should be amended to expressly recognise vulnerabilities in different population groups and particularly the gendered nature of some harms. The recognition of particular sub-groups in assessing harm (cl 45(4) and cl 46(4)) does not go far enough.

An important factor in balancing rights is Article 8 ECHR which requires a State to protect individuals' physical and psychological integrity; the case law suggests that where there are no protections against speech aimed at attacking or denigrating people, a state is in violation of its Convention obligations (see e.g. *Beizaras and Levickas* v. *Lithuania*44 ). A balance must be achieved between the two rights. The OSB should be in the spirit of this approach.

**Content in Scope**

**The draft Bill specifically includes CSEA and terrorism content and activity as priority illegal content. Are there other types of illegal content that could or should be prioritised in the Bill?**

Fraud (which is currently not in scope at all), hate speech that meets the criminal threshold (including misogyny, although it is not currently a specific crime), threats to harm, injure or kill, harassment.

On fraud, we have been working with a coalition of consumer, industry and charities to make the case that, without the inclusion of online fraud, there is a risk of complex and

41https://www.gov.uk/government/publications/intimidation-in-public-life-a-review-by-the-committee-on-standards-in-public-life

42 https://www.ipinst.org/2021/06/protection-related-barriers-to-womens-participation-in-northern-ireland-paper

43See also forthcoming Reset survey on self-censorship

44 https://hudoc.echr.coe.int/eng#{%22itemid%22:[%22001-200344%22]}

muddled regulations, and far worse consumer outcomes than an Online Safety Bill with a comprehensive approach to online fraud.

While we welcome the recent inclusion in the Bill of fraud carried out through user-generated content and fake profiles on social media websites, there is still a long way to go. Failing to include online advertising in the Bill leaves too much room for criminals to exploit online systems.

This view is backed by the FCA[45], Bank of England[46], City of London Police[47], Work and Pensions Committee and Treasury Committee[48], who have all commented that the scope of the Online Safety Bill should be expanded to include fraud carried out via online advertising. As we proposed in a previous blog post, designing in a framework for "interlocking regulation" would enable the expansion of scope without over-burdening OFCOM.[49]

*Action: the Government must include online advertising in the scope of the OSB.*

**The draft Bill specifically places a duty on providers to protect democratic content, and content of journalistic importance. What is your view of these measures and their likely effectiveness?**

We agree that there should not be double regulation of the traditional media sectors, and that political speech is deserving of a high level of protection. We are aware that some bodies with specialist expertise in these areas will probably comment in detail[50] so we make some initial, brief points mainly reflecting the areas where the lack of information provided by the Government need to be addressed:

> (1) what is democratic content (especially the meaning of cl 13(6)(b)), and would it exclude material that would fall foul of Article 17 ECHR[51] on using Convention rights to abuse the rights of others?;

45 https://committees.parliament.uk/oralevidence/2155/html/

46 https://www.ft.com/content/aa0f0763-8692-4211-92e0-c9bcb2655d0e

47 https://news.cityoflondon.gov.uk/urgent-action-needed-on-fraud-warns-city-of-london-police-authority-board/

48 https://committees.parliament.uk/committee/158/treasury-committee/news/156885/online-safety-bill-committees-warn-prime-minister-over-lack-of-action-on-harmful-paidfor-scam-adverts/

49 https://www.carnegieuktrust.org.uk/blog-posts/online-harms-interlocking-regulation/

50 See submission to the Joint Committee from the following organisations: Antisemitism Policy Trust, Center For Countering Digital Hate, Clean Up The Internet, Compassion In Politics, Demos, Elect Her, Fair Vote UK, Glitch, HOPE Not Hate, IMPRESS, Institute For Strategic Dialogue, Reset, SumOfUs, Unlock Democracy and Who Targets Me. Found here: https://drive.google.com/file/d/1MgqAHQo9BaLKl8gLDFl2fWCE20EK6pcm/view

51 See generally 'Guide on Article 17 of the European Convention on Human Rights- European Court of Human Rights https://www.echr.coe.int/Documents/Guide_Art_17_ENG.pdf

(2) does it cover illegal democratic content (e.g. incitement to riot in support of a political cause, or to target immigrants/refugees, or plain hate speech)?;

(3) what about e.g. covid disinformation/climate change denial and other forms of disinformation (where people can suffer harm as a result)?

(4) the definition of journalism both too wide (with no quality threshold) and too narrow (in that it excludes some modern forms of journalistic content);

(5) the provisions do not cover abuse of journalists, nor do they consider when journalistic privilege is exploited to abuse others; and

(6) what is relationship between these provisions and the general freedom of expression and privacy obligations?

We note that the International Peace Institute Report on Women's Participation in Northern Ireland of June 2021, citing the Gillen Review noted that

> *"There are also signs of the use of "journalistic privilege" to defend the comments made on public figures on social media, which contributes to a climate of impunity for online harassment and abuse."*[52]

In sum, as drafted, we have concerns that these provisions are capable of exploitation, undermining the effectiveness of the regime as a whole, while not necessarily protecting widely enough.

*Action: The Government needs to provide much more clarity in relation to democratic content and content of journalistic importance*.

**Earlier proposals included content such as misinformation/disinformation that could lead to societal harm in scope of the Bill. These types of content have since been removed. What do you think of this decision?**

We disagree with the decision to exclude misinformation/disinformation from the sorts of harms which the platforms should take into account when mitigating harm.

There seems to be a disconnect (deliberate or accidental) between the UK security establishment, who talk up the threat from disinformation and misinformation, and the draft OSB which largely ignores it, despite regulating primary channels through which disinformation flows. Civil regulation supports national security, with democratic oversight, in other regulated sectors and should do so here.[53]

---

52"At the Nexus of Participation and Protection: Protection-Related Barriers to Women's Participation in Northern Ireland"; CATHERINE TURNER AND AISLING SWAINE (June 2021) : https://www.ipinst.org/2021/06/protection-related-barriers-to-womens-participation-in-northern-ireland-paper

53 Stewart McDonald MP, "Disinformation in Scottish Public Life" (https://www.stewartmcdonald.scot/files/disinformation-in-scottish-public-life-june-2021.pdf); APPG on Technology and National Security, "How can technology increase the UK's resilience to misinformation during the next General Election?" (https://www.appgtechnatsec.com/resources)

The Director General of the Security Service referred repeatedly to the threat to the UK arising from disinformation and misinformation in his 2021 Threat Assessment[54] and said:

> "*we need a whole-of-system response, joining up not only across Government but also going much wider into industry and academia, and sometimes through to individuals.*"

The draft OSB does not meet the Director General's requirement. It should be an opportunity to lock significant platforms into a risk assessment mechanism for threats to security from mis- and disinformation under regulatory supervision, with appropriate transparency to Parliament. The draft OSB could also formalise and make more transparent the manner in which the UK public sector communicates threat assessment to platforms through the operation of the Counter Disinformation Cell in DCMS[55]. The Cell should be put on a formal statutory footing with an obligation to report to Parliament and to include OFCOM in its work.

On societal harms more broadly, we are concerned that the limitation of harms to individuals will not help the regime tackle issues such as high levels of misogyny and racism on a service which might undermine social cohesion, and indeed then feed back into harms to individuals. An avalanche of hateful speech in a public forum may have a greater effect on society than the sum of harms to individuals against whom it is directed. It appears that the Secretary of State can address such issues through their "Priority Harms" which are not constrained by the limitation to individuals. The Secretary of State should make an indication of "Priority Content" by the end of 2021 to assist scrutiny (although see our comments below on reforming the "Priority Content" approach).

The proliferation of misinformation and disinformation also has a corrosive effect on the country's "epistemic security",[56] on people's ability to access and identify reliable information across a range of issues. There is, moreover, a concern that untargeted scepticism undermines the ability to persuade on the basis of sound evidence, when all information is presented as qualitatively equal. Media literacy can help deal with some aspects of this problem but on its own is insufficient and risks shifting the responsibility back onto the individual user rather than the system.

*Action: OSB needs to connect to national security apparatus and regulate for national security with democratic oversight. The Counter Disinformation Cell should be put on a formal statutory footing with an obligation to report to Parliament and to include OFCOM in its work. The Secretary of State should indicate Priority Harms that address societal harm such as racism.*

---

54 https://www.mi5.gov.uk/news/director-general-ken-mccallum-gives-annual-threat-update-2021

55 Caroline Dinenage letter to Lord Puttnam, 29 May 2020 https://committees.parliament.uk/publications/1280/documents/11300/default/

56 E. Seger, S Avin, G Pearson, M Briers, S. OhÉigeartaigh Tackling threats to informed decision-making in democratic societies: Promoting epistemic security in a technologically-advanced world, 14 October 2020, https://www.cser.ac.uk/resources/epistemic-security/

**Are there any types of content omitted from the scope of the Bill that you consider significant e.g. commercial pornography or the promotion of financial scams? How should they be covered if so?**

We refer to our comments on scams and misinformation above. One significant omission is the exclusion of advertising. The concern is not necessarily about the content of advertising, but the question of whether the systems that drive advertising would be caught by the regime, if advertising is not. This is important because the advertising aspect of these platforms is a significant driver of harms (see evidence from Centre for Countering Digital Hate)[57].

**What would be a suitable threshold for significant physical or psychological harm, and what would be a suitable way for service providers to determine whether this threshold had been met?**

The threshold of psychological or physical harm is significant – if this is too high then this part of the regime will be greatly limited in its effect. Note the requirement is that the adverse impact must be "significant". The meaning of "psychological harm" is potentially problematic in this regard. Given the regime is based on the duty of care, existing meanings from tort law may affect the threshold. In tort law, similar-sounding thresholds for psychological harm have been set so high as to be of little use. They tend to revert to something like "a recognised psychiatric condition/injury" i.e. a medical definition. Similar concerns arise in the criminal law context – the Law Commission has criticised both.[58]

We also note the proposals from the Law Commission regarding communications offences[59] . Given the severity of criminal sanctions, it seems to us logical that the threshold for regulatory intervention should be lower than that for criminal penalties. Moreover, the Law Commission's proposal on the revision of s.127 Communications Act to tighten the conditions for the criminal offence is predicated on the assumption that the regulatory regime will be dealing with content creating a lesser level of harm. The thresholds should be at least the same as those currently applicable to video-sharing platforms in the Communications Act[60] (which affects a subset of operators that will be covered by the online safety regime).

More specifically, the draft OSB is not clear as to whether an assessment of harm is to be done by considering the impact of an individual item of content, or the cumulative impact of such content taken together (note that the word content is the same whether referring to either a single item or to multiple items) or across multiple platforms. The case of the abuse directed towards the England footballers is a case in point. While some examples would reach the criminal threshold, it is far from clear that all would (e.g. instances of monkey or banana emojis), yet the cumulative impact is great.

57 https://committees.parliament.uk/writtenevidence/38805/html/

58 Law Commission Liability for Psychiatric Illness, 10 March 1998 (LC249); Law Commission, Harmful Online Communications: The Criminal Offences, 11 September 2020 (Consultation Paper 248).

59 https://www.lawcom.gov.uk/project/reform-of-the-communications-offences/

60 Ss 368Z1and 368E Communications Act

A similar point could be made in relation to self-harm and suicide information (that does not meet the threshold for glorification, which is the requirement for the proposed new criminal offence). A person who searches for that sort of information, thus triggering the repeat delivery of it due to personalisation systems, might be peculiarly vulnerable to being influenced by it yet if the assessment of harm is made on the basis of each item of content individually, might fall outside the regime. This example is one where potentially neutral content is transformed into a hazard by the operation of the system; tackling the underlying content directly would be disproportionate when the issue is the personalisation (over which the speaker has no control).

Rare insights into company processes such as the New York Times reporting[61] on the "Bad for the world" experiment suggest that the companies are aware of risks but chose not to act through flawed harm reduction processes that may prioritise commercial interest over consumer harm. This is one reason increased transparency is so important.

OFCOM is well practised in research that could help determine such thresholds and could be asked by the Secretary of State (or by the Committee) to produce a paper to inform scrutiny.

*Action: OFCOM to be asked - by the DCMS Secretary of State or by the Joint Committee - to produce a paper to inform the scrutiny process with regard to the thresholds for physical or psychological harm in the regime.*

**Are the definitions in the draft Bill suitable for service providers to accurately identify and reduce the presence of legal but harmful content, whilst preserving the presence of legitimate content?**

In the UK, different types of media has been regulated for harm in different ways for decades. Current media regulation prohibits content that is harmful[62], leaving the regulator to give more detailed guidance as to what that means; and then companies make judgements about compliance. The courts have been quite content with OFCOM's process and guidance, as can be seen in the failed attempt by the Free Speech Union to judicially review OFCOM's decisions on COVID19 issues.[63] Similarly, advertising regulation prohibits harmful content and, moreover, prohibits the advertising of certain products. Social media companies are world-leading experts about people's reaction to the media the company's systems and process choose to display to them. They will be capable of working out what is harmful and are likely already to know that.

Media regulators (OFCOM and its predecessors) and media self-regulatory bodies (BBFC, ASA) in the UK have a decades-long track record of qualitative and quantitative research

61 'Facebook Struggles to Balance Civility and Growth' 24 November 2020 https://www.nytimes.com/2020/11/24/technology/facebook-election-misinformation.html

62 S 319 Communications Act 2003

63 Free Speech Union and Toby Young v OFCOM [2020] EWHC 3390 (Admin): https://www.bailii.org/ew/cases/EWHC/Admin/2020/3390.html

into the impact of media upon people to carry out these duties. Given this experience, it seems to us that the regulator and industry should be perfectly capable of "filling in" the detail of the harms caught by the regime[64]. The regime could work as follows: OFCOM should work with the social media industry to understand people's expectations of these thresholds informed in particular by the experience of victims and reflect that in codes of practice (which may also improve the inclusiveness of the online environment); service providers should use their formidable research powers to understand their customers experience and act to reduce harm.

To meet the Secretary of State's objectives, OFCOM's clause 61 review of harm should be as wide-ranging as possible. This outcome of this could well challenge the artificial multi-part characterisation of harm (children, illegal, adults, priority etc). In a Bill which is (deliberately) fragmented, the clause 61 review provides a rare moment of coherence.

As we said above these issues would be simplified by the introduction of an overarching duty of care. We shall bring forward amendments to implement such a duty in due course.

*Action: reinstate a single overarching duty of care.*

**Services in Scope**

**The draft Bill applies to providers of user-to-user services and search services. Will this achieve the Government's policy aims? Should other types of services be included in the scope of the Bill?**

Ed-tech is out of scope but there should be an evaluation of the experience of the sudden shift to online learning during the 2020-21 lockdowns to understand whether ed-tech services - which can demonstrate many of same criteria/functionality as social media platforms, e.g. data collection, communication functionality etc - should be in scope.[65]

*Action: OFCOM designate the Children's Commissioner to review the evidence on ed-tech as part of Clause 61 review.*

**The draft Bill sets a threshold for services to be designated as 'Category 1' services. What threshold would be suitable for this?**

---

64 This point was made by Baroness Grender in a Lords debate on Social Media Services in 2018: "If in 2003, there was general acceptance relating to content of programmes for television and radio, protecting the public from offensive and harmful material, why have those definitions changed, or what makes them undeliverable now? Why did we understand what we meant by "harm" in 2003 but appear to ask what it is today" https://hansard.parliament.uk/Lords/2018-11-12/debates/DF630121-FFEF-49D5-B812-3ABBE43371FA/SocialMediaServices

65 See Carnegie UK "Closing the Divide" report: https://d1ssu070pg2v9i.cloudfront.net/pex/pex_carnegie2021/2021/06/05091733/CUKT-Time-for-Action-Advocacy-FINAL.pdf

The threshold at present is a combination of size and risk. In a truly risk-managed regime, there should be the possibility for some services to be Category 1 on risk alone. We have in mind the example of the former Canadian service, Kik, which (although small in the UK) was reportedly[66] the subject of 1,100 reports of child abuse to England and Wales police. OFCOM defining the threshold based on evidence is an appropriate route, but OFCOM could be asked to provide a provisional answer without prejudice during scrutiny to assist deliberation.

*Action: Secretary of State to ask OFCOM to give without prejudice estimate of threshold by end of 2021 to assist scrutiny.*

**Are the distinctions between categories of services appropriate, and do they reliably reflect their ability to cause harm?**

The Government has not provided sufficient explanation of the lesser responsibilities imposed on search engines which lead to a significant gap in protection.

We note that the transparency obligations will be determined by reference to further categorisations (Cat 2A and 2B). The scope of these categories is unclear. The scope of the transparency obligations are, in our view, unnecessarily limited as we have argued above.

*Action: Secretary of State to provide more detail on justification for search engine exemptions.*

*Secretary of State to ask OFCOM to give without prejudice estimate of Category 2A and 2B thresholds by end of 2021 to assist scrutiny.*

**Will the regulatory approach in the Bill affect competition between different sizes and types of services?**

The Competition and Markets Authority has demonstrated that the digital advertising market is distorted in favour of the largest operators, two of which will be covered by the OSB[67]. The OSB will return back to the companies some of the external costs they have shed onto society. That should make the market work more efficiently as long as the burden of regulation does not fall disproportionately on smaller service providers.

The OSB will bite hardest on Category One operators. For smaller operators, proportionality to their capability and risk is important. Increased choice might well allow a better match of user risk appetite to service provision, rather than the one-size-fits-all approach seen in dominant platforms, leading to reduced harm.

The Government seems to have chosen "proportionate" over "reasonable": the word used historically in defining duties of care and around which much case law exists. The Bill uses the word "proportionate" over two dozen times but does not fully define it (see in relation to the safety duties cl 9(6) and 10(6) but which relate only to those duties). It could mean

66 Kik chat app 'involved in 1,100 child abuse cases' Angus Crawford BBC News 21 September 2018 https://www.bbc.co.uk/news/uk-45568276 BBC News

67 https://www.gov.uk/cma-cases/online-platforms-and-digital-advertising-market-study

proportionate to company size, or risk or both: it does not consider the appropriateness or effectiveness of the measure. The word "proportionate" is linked to attempts early in the coalition government to reform economic (not safety) regulation.

> *Efficient and proportionate regulation*
>
> *Economic regulation, as with most forms of regulation, imposes costs on regulated companies. These costs derive from the regulatory cost the regulators impose on their sectors and the administrative cost of running the regulatory institutions. Costs in these sectors tend to be passed through to end consumers. It is important that they are proportionate and outweighed by the benefits achieved for consumers. Cost minimisation might, however, not always be efficient, as lowering costs can sometimes lead to foregoing bigger benefits to consumers.* (Principles of Economic Regulation BIS 2011)[68]

*Action: Secretary of State to provide more information on interpretation of "proportionate", such as reference to case law from other areas of regulation.*

**Algorithms and user agency**

**What role do algorithms currently play in influencing the presence of certain types of content online and how it is disseminated? What role might they play in reducing the presence of illegal and/or harmful content?**

Most social media services have not been neutral conduits with regard to content for a long time. We understand that companies continually tweak their models to maximise corporate objectives. An example is Facebook's 2020 "Bad for the World" experiment*:*

> *"The team trained a machine-learning algorithm to predict posts that users would consider "bad for the world" and demote them in news feeds. In early tests, the new algorithm successfully reduced the visibility of objectionable content. But it also lowered the number of times users opened Facebook, an internal metric known as "sessions" that executives monitor closely.*
>
> *"The results were good except that it led to a decrease in sessions, which motivated us to try a different approach," according to a summary of the results, which was posted to Facebook's internal network and reviewed by The [New York] Times*."[69]

Regulation of the synthetic environments created by social media companies should seek to rebalance the system in a pro-social manner with independent, evidence-based oversight structurally similar to the way broadcasting, advertising and film are regulated, rather than playing "whack a mole" with individual pieces of content which will not work at scale.

---

68 BIS URN 11/795 Found at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/31623/11-795-principles-for-economic-regulation.pdf

69 See https://www.nytimes.com/2020/11/24/technology/facebook-election-misinformation.html

Carnegie worked with representatives of victims of hate crime to develop a hate speech code of practice (attached in annex B) which considers algorithmic issues - see answer to following question.

**Are there any foreseeable problems that could arise if service providers increased their use of algorithms to fulfil their safety duties? How might the draft Bill address them?**

The use of automated tools does give rise to risks. Platforms should also be aware of the biases inherent in their datasets, as well as in their design choices and definition of the parameters and outcomes of the system. Currently there is evidence, for example, that automated moderation tools exhibit bias in the content flagged for deletion.

Carnegie UK worked with representatives of victims of hate crime to produce a draft code of practice on hate crime and social media (mentioned in oral evidence by Danny Stone of the Antisemitism Policy Trust[70]). The code of practice (attached in annex B) set out a pragmatic approach to addressing hate speech which balances with expert human input the evident failings of pure algorithmic approaches to meeting safety duties. We commend it to the committee.[71]

The code also cites the existing law, clarified by the Government in response to Lord Stevenson of Balmacara that algorithms should be tested for safety before deployment in a workplace[72].

*Action: Government to adopt the Draft Hate Crime Code of Practice drafted by Carnegie UK and other civil society organisations.*

**Does the draft Bill give sufficient consideration to the role of user agency in promoting online safety?**

In responding to this question, one has to be wary of victim-blaming. A person of "reasonable sensibilities" might be deterred from joining a social network called 'Horrificabuse.com' unless they had a penchant for such material. In this example the risk of harm is clear. There is little indication however of the potential for harm to occur on the mainstream networks that might fall into Category One. Shifting the responsibility for preventing harm onto the victim in general purpose media networks is not how the UK has regulated other mass-market media platforms; while the move to 'pull services' (video on demand) indicated greater user choice, certain minimum standards were required and users were to be provided with information on which to base their choices. Traditional broadcasting (which pushes content at its audience) was subject to stronger constraints (and it might be argued that services that have autoplay features default on might also be push services).

70 https://committees.parliament.uk/oralevidence/2695/html/

71 See code here: https://www.carnegieuktrust.org.uk/publications/draft-code-of-practice-in-respect-of-hate-crime-and-wider-legal-harms-covering-paper-june-2021/

72 UIN HL8200, tabled on 23 May 2018 https://questions-statements.parliament.uk/written-questions/detail/2018-05-23/HL8200

When design and process-based methods of preventing harm fail or while they are being improved, it seems proportionate to give people easy to use "self-defence" tools. Prof Woods and William Perrin worked with Anna Turley MP as long ago as 2016 on such an approach in her Malicious Communications (Social Media) Bill 2016-17[73]. The Bill required that platforms only provide a stream from which threatening content has been filtered unless a user aged over 18 explicitly requests the unfiltered version. We note that user defence tools have improved in recent years, which is welcome. However user-defence tools do little to assist people who are harmed but are not members of that platform and more must be done to consider the problems upstream.

*Action: OFCOM should include adequacy of user defence tools as part of a risk assessment.*

**The role of Ofcom**

**Is Ofcom suitable for and capable of undertaking the role proposed for it in the draft Bill?**

Yes. We have long called for OFCOM to take on this role. As a regulator, OFCOM has proven itself capable of changing to take on new roles several times. OFCOM staff and board have huge experience of the processes necessary to arrive at decisions on subjective matters in the public interest; and then defending those decisions in the courts. We have been encouraged by OFCOM making many hires of high-quality people from expert civil society organisations to deliver its new functions.

**Are Ofcom's powers under the Bill proportionate, whilst remaining sufficient to allow it to carry out its regulatory role? Does Ofcom have sufficient resources to support these powers?**

The draft Bill provides for the first time a mechanism by which a regulator has the right combination of tools (powers to request information, require risk assessments, sanction and disrupt business etc) to hold global social media companies to account. If one abstracts this out, the powers that will drive regulatory compliance could with little adaptation be used to tackle other regulatory problems arising from social media. This is a substantial achievement by DCMS. Regulators (such as the FCA, Trading Standards) could use this tool set to address issues that arise outside the immediate focus of online safety. This gives rise to two issues:

- how do regulators co-operate to achieve social objectives with OFCOM? For instance the FCA taking a dossier to OFCOM on harm caused by social media in financial services that was triggered by a systems failure (e.g poor "Know Your Customer" processes) that needs to interlock with, say, financial services?
- how are powers delegated from OFCOM to other bodies (sometimes called "co-designation") such as to the Internet Watch Foundation)?

73 Malicious Communications (Social Media) Bill (HC Bill 44)

https://publications.parliament.uk/pa/bills/cbill/2016-2017/0044/cbill_2016-20170044_en_2.htm#l1g1

We discuss in our blog on interlocking regulation how regulators might cooperate[74]. In terms of co-designation, OFCOM could use the powers to delegate functions under the Deregulation and Contracting Out Act 1994 [75]in conjunction with s 1(7) Communications Act. In our view, however, it would be preferable to include a specific provision to allow delegation, which includes safeguards as to the bodies to which the powers could be designated. Models for such a clause already exist in the Communications Act (in relation to the designation of the ASA as regulator of advertising for on-demand programme services).

The bedrock of any regulatory regime is the regulator's ability to find out what is going on in the regulated companies. In the OSB, an important part of this is transparency reporting (clauses 49 and 50) which will allow observers and the regulator to track progress over time. In contrast to the framework nature of the rest of the Bill, the content of transparency reports is tightly drawn. There is no scope for OFCOM to add things in as its research reveals new problems, curiously, only the Secretary of State can add things. OFCOM would have to resort to making information requests under clause 70, etc, which seems more burdensome and also problematic in longitudinal tracking. Clause 49(5) provides sufficient safeguards to prevent OFCOM creating an undue burden if its power to influence transparency reports is increased. We noted above that only some companies are subject to transparency reporting; it is unclear where the threshold would be drawn. We believe this limitation should be removed, subject perhaps only to a *de minimis* exception.

*Action: Amend clause 49(4): replace "OFCOM may only describe information of" with "OFCOM may describe such information as it sees fit to execute its duties, including"*

*Action: remove limitation of transparency reporting in cl 49(3) (and delete 'relevant' in cl 49(1))*

For the largest or riskiest service providers it seems a strange omission that they should not have to share their risk assessments with OFCOM automatically, rather than OFCOM have to use an information request under clause 70 to obtain them. We also discuss in answer to a question above why a version of the risk assessments should also be published for the general public (discussed above).

*Action: In each of Clauses 9, 10 & 11 'Safety Duties' insert new final subclause "A duty to provide the risk assessment to OFCOM when it has been carried out as described in Clause 8."*

**How will Ofcom interact with the police in relation to illegal content, and do the police have the necessary resources (including knowledge and skills) for enforcement online?**

74 https://www.carnegieuktrust.org.uk/blog-posts/online-harms-interlocking-regulation/

75 https://www.legislation.gov.uk/ukpga/1994/40/section/70

The have been under resourced in this regard since online media began. While the police may well require greater resources and more training, which is another cost that the social media companies would be pushing on to society, there are circumstances where criminalising speech is not appropriate (eg where speech is turned into a hazard through the operation of the platform).

That is why a civil regulatory regime with an expert regulator acting in concert with a criminal regime is necessary to protect people from the external costs of badly run social media. The Law Commission, in their report on Communications Offences, acknowledged the need for a civil regulatory regime and that the criminal regime was a backstop to that, with higher thresholds for harm.

*Action: the Joint Committee should take evidence from the National Police Chiefs Council lead on hate speech/digital issues*

**Are there systems in place to promote, transparency, accountability, and independence of the independent regulator?**

The United Kingdom is party to Recommendation Res(2000)23 of the Committee of Ministers of the Council of Europe on the independence and functions of regulatory authorities for the broadcasting sector. The Recommendation says that:

> *The rules governing regulatory authorities for the broadcasting sector, especially their membership, are a key element of their independence. Therefore, they should be defined so as to protect them against any interference, in particular by political forces or economic interests....For this purpose, specific rules should be defined as regards incompatibilities in order to avoid that:*
>
> *- regulatory authorities are under the influence of political power;*
>
> *- enterprises or other organisations in the media or related sectors, which might lead to a conflict of interest in connection with membership of the regulatory authority.*
>
> *Furthermore, rules should guarantee that the members of these authorities:*
>
> *- are appointed in a democratic and transparent manner;*
>
> *- may not receive any mandate or take any instructions from any person or body;* [76]

We are concerned that the draft OSB contravenes the general principle of Res(2000)23. The Government has not explained why the Secretary of State needs these powers. We propose amending these powers to create a more conventional balance between democratic oversight and regulatory independence to underpin freedom of expression. As described in answer to previous questions, we have set this out in our recent blog article. The points made earlier in this submission as well as in the blog are relevant here.[77]

76 https://rm.coe.int/16804e0322.See

We shall provide detailed amendments to the draft Bill to give this effect in due course.

We also note that the competition for Chair of OFCOM is to be re-run despite candidates qualifying for appointment under the initial rules. Whilst the Government is following due process in re-running the competition this does not sit easily with the UK's international commitments in Res(2000)23. We note also the Comments from the Chair of the Commons DCMS Committee about the appointment.

*Action: Achieve a better balance between Parliament and the Secretary of State to preserve OFCOM's independence (as described in Carnegie blog post) and amend the Bill accordingly; this would include deleting Clause 33 and Clause 113.*

**How much influence will a) Parliament and b) The Secretary of State have on Ofcom, and is this appropriate?**

Parliament's day-to-day relationship with OFCOM seems broadly similar in the new regime once the initial SIs have been made. There are a lot of SIs: careful consideration will be required as to whether each has been given the appropriate level of Parliamentary oversight.

As we discuss in our blog post referred to above, the Secretary of State's power to direct OFCOM to change codes of practice to bring them in line with Government policy compromises OFCOM's independence (see answer above). It is noteworthy that the exercise of some of those powers (cl 33 and 113) are not subject even to Parliamentary oversight. As noted, we propose the deletion of the two provisions; the priority harms processes and the strategic direction of OFCOM will necessarily involve the executive in the process but the exercise of these powers should be subject to close Parliamentary scrutiny.

A better balance can be struck between Parliament and the executive in setting priorities that maintain OFCOM's independence. We suggest examining the issue in two parts: regime start up; and response to issues during operation. The draft Bill should be amended so that:

- the Secretary of State specifies (with supporting research) the initial outcomes they seek to address and 'priority content' on the face of the Bill, which Parliament can hold to account. This sets priorities during the regime start-up phase.
- during regime operation, changes to priority content should originate from OFCOM's research, not from the Secretary of State, and be rigorously evidence-based. OFCOM should form the need for new priority content from its research, then consult Parliament, the Secretary of State and others. OFCOM should have regard to the consultation and present a report to the Secretary of State from which they should make a Statutory Instrument (by the positive procedure) to put the new priority content into effect.

The Secretary of State should periodically (every three years) be able to give OFCOM an indication of their strategic priorities for Internet Safety, but this should not cut across into

77 https://www.carnegieuktrust.org.uk/blog-posts/secretary-of-states-powers-and-the-draft-online-safety-bill/ and also at annex A

content, nor into OFCOM's day-to-day administration, and should remains subject to Parliamentary oversight.

We refer to our earlier comments on clauses 33 and 113.

The Bill gives powers to the Government to act in an emergency to require the publication of 'public statement notices' (Clause 112). We do not understand what outcome this Clause is intended to deliver in a crisis situation and it requires further explanation.

*Action: Delete clause 33 and clause 113. Amend cl 109 to ensure that the strategic priorities remain high level objectives, not control over day-to-day implementation of the regime. The Government should also explain or modify Clause 112.*

**Does the draft Bill make appropriate provisions for the relationship between Ofcom and Parliament?**

Parliament in many regulatory regimes has been content to set a high-level framework and delegate to the regulator the day-to-day running of a regime. This gives the regulator flexibility to respond and adapt to changing circumstances. The regulator is fleet of foot in a way Parliament could never be, giving regulated companies and consumers the confidence that decisions will be taken. Where the regulator is under resourced or leashed too tightly to a rigid rule-set dominant companies with huge resources can and will game and frustrate the regime. The first ten years of OFTEL's existence are an example. In such circumstances, Government and Parliament have to legislate frequently to keep up - this creates considerable hazard in the regime and further favours dominant operators.

The British model of an independent regulator acting independently of the Government to implement Parliament's wishes in a complex world is highly distinctive. Lord Currie discusses the merits in his 2014 lecture78. A delegated regime is highly appropriate in areas dominated by technological change and fast-moving problems. Parliament has for decades trusted independent, competent media regulators to determine difficult issues, as Lord Currie noted:

> *"It has long been accepted that there is content that is within the law but the dissemination of which should be controlled: in broadcast media, regulators (now Ofcom) have enforced standards for accuracy and impartiality and to avoid harm and offence, though with increasing difficulty in a world of proliferating channels, standards that are over and above what the law requires. Given the sensitivity of these issues, government has preferred this to be done independently and at a distance."*

The draft OSB is a framework with much to be filled in through secondary legislation. We note that only the Secretary of State's first round of priority harms has to follow a positive resolution procedure. Subsequent rounds will only require the negative procedure giving

78 Lord Currie: "'The case for the British model of independent regulation 30 years on" Cass Business School (21 May 2014) https://www.gov.uk/government/speeches/the-case-for-the-british-model-of-independent-regulation-30-years-on

the executive too much power on matters of free expression (see our proposals for reform set out in our blog post on the Secretary of State's powers and referred to in answers above). We note the Lords Communications Committee suggestion of a Joint Committee to consider secondary OSB legislation, which has much merit[79].

**Is the status given to the Codes of Practice and minimum standards required under the draft Bill and are the provisions for scrutiny of these appropriate?**

We suggest removing the Secretary of State's power to interfere with these codes, give Parliament more influence at the outset and more flexibility for the regulator downstream - see previous answer.

**Are the media literacy duties given to Ofcom in the draft Bill sufficient?**

OFCOM's 2021 Adults' Media Use and Attitudes survey shows there is still a long way to go on media literacy, particularly amongst lowest socio-economic groups, with only 53% in social groups DE aware that some websites will be biased/inaccurate.[80]

A simple segmentation would be into people (still) using platforms and people who don't.

For the first group, OFCOM should use the power of the regulated platforms to improve media literacy but under objective supervision rather than at the platforms whim. One route would be to mimic the approach in the rest of the Bill in respect of harms. For instance, OFCOM could identify in its clause 61 survey of harms those which could best be mitigated by media literacy. OFCOM could then require regulated platforms to report as part of their risk assessments and harm reduction plans how they will use media literacy as a tool to reduce harm and set out a plan with clear deliverables. The platforms also provide details of evaluation of effectiveness of those measures, informed by research agreed with OFCOM and publish the results of that evaluation. This introduces an element of economically-efficient “polluter pays”. It isn't clear if Clause 103 gives OFCOM the power to do this.

However the second group - who may have been deterred by online safety concerns (whether as victims or are simply scared) - are much harder to reach. Media literacy and digital literacy go hand in hand. In particular, people who are not in education or training and have few connections to educational institutions. Those with lower digital engagement are less likely to be media savvy:

- Narrow internet users (who complete fewer activities online) generally had a lower than average critical understanding of the online environment (Ofcom 2021).
- Still a large number of people in this space - 14.9 million adults currently have "very low digital engagement"[81].

79 https://publications.parliament.uk/pa/ld5802/ldselect/ldcomuni/54/5402.htm

80 https://www.ofcom.org.uk/research-and-data/media-literacy-research/adults/adults-media-use-and-attitudes

81 https://www.lloydsbank.com/banking-with-us/whats-happening/consumer-digital-index.html

Post pandemic, people's views on where they go for help with media literacy has changed. In 2021, there has been so far a seven-fold increase (35% compared to 5% in 2020) in the number of people wanting somewhere local to build their digital confidence and skills (Lloyds 2021). OFCOM should encourage regulated companies to consider community-based organisations, such as Good Things Foundation, in their measures to tackle digital literacy among this group. (Declaration - William Perrin is also a Trustee of Good Things Foundation)

The Government has stopped funding the Future Digital Inclusion programme. This has left very excluded adults for whom formal education isn't a familiar or trusted route to either be left behind or seek support from necessarily much smaller philanthropically funded digital inclusion programmes.

The Essential Digital Skills framework could be reviewed to see if it is sufficient for avoiding harms. The particular duties on disinformation (Clause 122) which derive from OFCOM's media literacy powers are quite odd and we do not understand how they are intended to work.

*Action: OFCOM should include media literacy issues in its Cl61 survey and follow and require media literacy to be built into risk assessments as a mitigation measure.*

*OFCOM should use its analytical ability to estimate how many people require media literacy training and the gap between that number and current capacity.*

*The Essential Digital Skills Framework should be reviewed to see if it is sufficient for avoiding harms.*

**Carnegie UK**

**September 2021**

**Contact: maeve.walsh@carnegieuk.org**

**ANNEX A: BLOG POST ON SECRETARY OF STATE POWERS (Published 14th September 2021)**

**https://www.carnegieuktrust.org.uk/blog-posts/secretary-of-states-powers-and-the-draft-online-safety-bill/**

**The draft Online Safety Bill gives too many powers to the Secretary of State over too many things.[1] This is a rare point of unity between safety campaigners, who want tough legislation to address hate crime, mis/dis-information and online abuse[2] and radical free speech campaigners who oppose much of the Bill.**

To meet the UK's international commitments on free speech in media regulation, the independence of the regulator from Government is fundamental. This boundary between the respective roles of the Government and the regulator in most Western democracies is well-established. The United Kingdom is party to a Council of Europe declaration that states that national rules for a broadcasting regulator should:

> "*Avoid that regulatory authorities are under the influence of political power.*"

The United Kingdom was also party to a 2013 joint statement on freedom of expression between the Organisation for Security and Co-operation in Europe (OSCE) (of which the UK is a participant), the Office of the United Nations High Commissioner on Human Rights, the Organisation of American States and the African Commission on Human and Peoples' Rights. In that statement, made at a time of great international regulatory change due to the move to digital transmission, the United Kingdom also agreed that:

> *"While key policy decisions regarding the digital terrestrial transition need to be taken by Government, implementation of those decisions is legitimate only if it is undertaken by a body which is protected against political, commercial and other forms of unwarranted interference, in accordance with international human rights standards (i.e. an independent regulator)."*

The United Kingdom has been a leading exemplar of the independent regulator approach. In the Communications Act 2003, Parliament set OFCOM a list of objectives for setting its standards codes, then leaves OFCOM to set the codes without further interference or even having to report back to Parliament. This is a good demonstration of the balance referred to in the OSCE statement. Parliament and government set high-level objectives in legislation then do not interfere in how the regulator does its day-to-day business.

With the Digital Economy Act 2017, Parliament agreed that Government could direct OFCOM, but that power was limited to exclude OFCOM's content rules. The Wireless Telegraphy Act 2006 powers of direction also do not touch content.

Unfortunately the draft Online Safety Bill deviates from these sound principles and allows the Secretary of State to interfere with OFCOM's independence on content matters in four principal areas. The draft Bill gives the Secretary of State relatively unconstrained powers to:

- set strategic priorities which OFCOM must take into account (cl 109 and cl 57)

- set priority content in relation to each of the safety duties (cl 41 and 47)

- direct OFCOM to make amendments to their codes to reflect Government policy (cl 33)

- give guidance to OFCOM on the exercise of their functions and powers (cl 113).

The UK Government has not explained why the Secretary of State needs these powers. We propose that the draft Online Safety Bill provisions relating to these powers should be amended to create a more conventional balance between democratic oversight and regulatory independence to underpin freedom of expression.

**Parliament and Government set OFCOM's initial priorities**

Parliament and Government, working with the traditional checks and balances, should be able to set broad priorities for OFCOM's work on preventing harm. We understand that OFCOM would also welcome initial prioritisation, as would regulated companies. Victims' groups also want reassurance the harms that oppress them will be covered by the legislation. Parliament will want to be confident in what OFCOM will do with the powers being delegated to it.

However, the Secretary of State's powers should not cross the line in the Digital Economy Act and permit the Government to direct OFCOM on content matters through Statutory Instruments (SIs). Clauses 109 and 57 do so on strategy (albeit with some Parliamentary oversight in cl 110) and cl 41 and cl 47 on Priority Content. These extensive powers enable detailed government influence on the implementation of policy, potentially influencing decisions that impact content, and undermine OFCOM's independence.

A better balance can be struck between Parliament and the executive in setting priorities that maintain OFCOM's independence. We suggest examining the issue in two parts: regime start up; and response to issues during operation. The draft Bill should be amended so that:

- the Secretary of State specifies (with supporting research) the initial outcomes they seek to address and 'priority content' on the face of the Bill, which Parliament can hold to account. This sets priorities during the regime start-up phase.

- during regime operation, changes to priority content should originate from OFCOM's research, not from the Secretary of State, and be rigorously evidence-based. OFCOM should form the need for new priority content from its research, then consult Parliament, the Secretary of State and others. OFCOM should have regard to the consultation and present a report to the Secretary of State from which they should make a Statutory Instrument (by the positive procedure) to put the new priority content into effect.

The Secretary of State should periodically (every three years) be able to give OFCOM an indication of their strategic priorities for Internet Safety, but this should not cut across into content, nor into OFCOM's day-to-day administration.

**Parliament and government then respect OFCOM's independence**

The draft Online Safety Bill envisages a continuing control in the hands of the Executive beyond high level strategic direction. Clauses 33 and 113 affect OFCOM's role to implement policy; the OSCE statement is particularly clear that this should be an area in which there is no Government interference. Yet both clauses cross the boundary emphatically. Moreover, there is no attempt to provide for scrutiny or control of these powers by Parliament. The Secretary of State's power to direct OFCOM to make amendments to the code to reflect Government policy (cl 33) and to give guidance as to the exercise of functions and powers are simply egregious and should be deleted.

**FOOTNOTES**

[1]

See table here: https://www.carnegieuktrust.org.uk/annex-b-the-role-of-the-secretary-of-state/ initial response

**[2]**

See Prof Lorna Woods quoted here: https://www.politico.eu/article/uk-concerns-over-internet-free-speech-tech-regulation-power-grab/

**ANNEX B:**

**DRAFT Code of Practice in respect of Hate Crime and wider legal harms: covering paper**

**June 2021**

**(Published Here: https://www.carnegieuktrust.org.uk/publications/draft-code-of-practice-in-respect-of-hate-crime-and-wider-legal-harms-covering-paper-june-2021/)**

**Introduction and background**

1. There are numerous codes of practice in place for social media companies. They are almost all voluntary in nature, limited in scope and have little to no impact in practice. Often, companies argue that their own activities go far beyond stated codes of practice which quickly b*ec*ome unfit for purpose. The initial consultation offerings from Government for the (then) Online Harms Bill indicated that a statutory Duty of Care would be introduced for social media platforms, or those allowing the sharing and distribution of user-generated content, and that Codes of Practice which informed the Duty of Care would be published.[82] A newly proposed regulator would, as one part of its wider examinations, assess adherence to the codes and judge failures to apply the Duty of Care accordingly. Codes on Terrorism and Child Sexual Exploitation and Abuse have been given prominence and special status, with oversight from the Home Secretary; drafts of these have since been published. The 'Hate Crime' Code of Practice has not been given such status and does not, as proposed, include wider harms.

2. That is why we have worked over the course of a number of months with a number of civil society organisations, who speak to the lived experience of many groups that experience hate crime online, on the development of this model Code of Practice for Hate Crime and wider legal harms.[83] A draft of this Code was discussed at a workshop in February 2021, which brought together those initial collaborators with representatives of the major platforms, policymakers and regulators. The headlines from their feedback are set out below and provide important framing for consideration of the Code itself as well as an indication of the different perspectives that we expect to encounter in this next stage of engagement and advocacy.

---

[82] The Online Harms White Paper (2019) provided examples of the codes of practice a regulator might draw up for companies to fulfil their duty of care in the following areas: CSEA and terrorist use of the internet (these voluntary codes, drafted by the government, were published with the Full Response to the White Paper consultation in December 2020); Serious Violence; Hate Crime; Harassment; Disinformation; Encouragement of self-harm and guidance; Online Abuse of public figures; Interference with legal proceedings; Cyberbullying; Children Accessing Inappropriate Content.
(https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf pp64-76)

[83] This draft Code has been developed with the invaluable input of the following organisations: Antisemitism Policy Trust, The Bishop of Oxford's Office, Glitch, Centenary Action Group, Faith Matters, Galop, Hope Not Hate, Institute for Strategic Dialogue, The Alan Turing Institute. It has also benefited from the insight and feedback on the draft Code from further contributors, including representatives of the major tech platforms, central government policymakers and regulators, at a workshop hosted by Carnegie UK Trust on 26th February 2021.

3. The Code draws on, and should be read in conjunction with, the Codes and other references in the annex, including the European Code of Conduct on Illegal Hate Speech Online and the Digital Economy Act Code of Practice. It offers systems-level solutions to addressing harms and could form the basis of a regulatory Code in this area. It is aimed at outlining principles rather than an exhaustive set of rules and service operators should aim to engage with the spirit of those principles and not look just to the letter of the Code. This draft Code does not at this stage synchronise fully with the Government's draft Online Safety Bill[84]. Furthermore, the work needs to be carefully considered in respect of devolved administrations and powers.

4. The challenges of discussing the merits and content of a draft Code without seeing the "parent" legislation to which it will be attached was a recurring theme at our February workshop. However, we did not wish to wait for that draft Bill in order to revise the Code "to fit" the Government's framework. With pre-legislative scrutiny about to start, we hope now to use this draft to start the debate on the need for the Government to include such a Code in the first place and also to re-emphasise the importance of a systemic (rather than content regulation) approach, from which a Code such as this could then naturally flow. While the draft Bill draws a boundary between the treatment of content that is likely contrary to the criminal law and that which is not, we still think there are merits in this holistic approach for two reasons. First, it demonstrates that not all instances of hate speech necessitate the same response and that in this we are not limited to take down. Secondly, it deals with the point that such speech can escalate and, crucially, minor-sounding instances can be linked to a larger picture of harm.

**Commentary on the draft Code**

5. Our workshop participants agreed that the strengths of the Code lay both in the content and in the means of its development. The systemic approach, with a focus on design and processes drawing on Carnegie's model[85], is positive. The Code builds on good practice but raises the bar, and the level of detail shows how such an approach would work effectively in practice. Keeping the emphasis on principles and outcomes will be important, as well as keeping the wording broad to take account of things that haven't yet emerged or not been thought of yet.

6. There is a "cart before horse" challenge (noted above) that arises because – at the time of our workshop – the Government's draft Bill had not yet been published at the time of our discussions. The development of the Code – ahead of the publication of the Government's legislative proposals – was seen as both a strength and risk. While it allows a truly systemic, process-led approach to be set out for debate and engagement, it was not clear how successfully it would integrate with the overarching legislation – and now, in the light of the draft Bill, that remains a challenge.

---

[84] https://www.gov.uk/government/publications/draft-online-safety-bill

[85] See, for example, our April 2019 reference paper (https://d1ssu070pg2v9i.cloudfront.net/pex/carnegie_uk_trust/2019/04/08091652/Online-harm-reduction-a-statutory-duty-of-care-and-regulator.pdf) and our draft Online Harm Reduction Bill (https://www.carnegieuktrust.org.uk/publications/draft-online-harm-bill/)

7. However, feedback suggested that the strength of this Code lay in the way it set out what could be seen as an overarching systems Code which could apply to other harms. (For example, one would not necessarily need the repetition of many of the clauses between multiple codes if there was an overarching systems code from which they flowed)**. One area for the Government and Parliamentarians to consider as it responds to scrutiny of the draft Bill is whether there is an opportunity to create a single core set of practices that would apply to all companies and then set addendums for specific harms and for specific companies**.

8. The balance between illegal activity and legal but harmful was judged to be about right. Workshop participants judged that it was important that the latter was included, particularly as there is a risk that the legislation may not give enough prominence to it; we are still working to understand the draft Bill on this aspect. The draft Code was also deemed to align well with international developments and best practice.

9. There were differences of opinion – inevitably – between civil society and tech company views on the necessary level of prescription and specificity in the Code. The former (in general) viewed the detail as a necessary baseline to raise the bar and ensure all companies comply, without introducing too much wriggle room for dilution of the requirements; the latter (representing the bigger platforms) would prefer less detail and greater flexibility to enable companies to develop solutions in a way that suits their specific circumstances.

10. While a concern was raised that the Code might treat all companies as the same, the emphasis on proportionality was welcomed and could be brought out further, particularly how it will work in practice. It was judged that a focus on proportionality will ensure that burdens on smaller companies will be minimised and, in the next stage of iteration, it will be important to consult more widely with them as well as the big platforms. That said, in terms of the detail and prescription, there was a view that the Code should not become too static: it needs to remain dynamic and flexible, to adapt to a rapidly changing landscape, and to remain focused on outcomes (which would remove some of the issues that companies might face if their particular systems and processes were not relevant for parts of the code).

11. Regardless of the systemic approach of this Code (and the opportunity it provides to set out a "template" for an overarching systems Code from which others might flow), the risk of duplication and a proliferation of codes is real issue: while the approach of the hate crime Code is complementary to those already published by the Government (on terrorist content and CSEA), it is not clear how they might fit together (or the respective status of codes, guidance, principles, etc) and plans to introduce further codes for individual harms may prove confusing and burdensome for companies, with competing requirements and different responses. There are also some boundary issues between this Code and the Code on terrorism.

<u>Specific areas of focus</u>

12. The **enforcement approach** was agreed in principle by participants at our workshop, but this needed to be balanced against safeguards. In particular, the proposals about data capture and passing data to law enforcement raised concerns, particularly around the requirement that platforms should make decisions about what content should be passed to criminal investigations and/or the government. The Code should align more closely with the existing processes that companies have for engagement with law enforcement agencies. The good Samaritan clause was also well balanced with due diligence.

13. The **information-gathering powers** are important – it is vital for civil society and regulators to understand better what is going on (eg banned accounts being set up again). Access to data is key to transparency between regulators and platforms. A focus on **metrics** was important (and would be headline grabbing) but it is important to get into the incentive models and the nuances around this. Too much specificity re metrics will mean that companies might game the system by focusing on numbers rather than why that particular thing is important, and could incentivise the worst behaviours.

14. The clauses on **trusted flaggers** were important: these roles bring more understanding of the communities affected by hate crime than the platforms but there needed to be more clarity on what these programmes should look like, labour and resource demands and the clarification of accountability and liability.

15. The collaborative, iterative approach facilitated by Carnegie UK Trust was welcome and was seen as a strength. Participants felt that Ofcom should be encouraged to set out how it will meaningfully engage with civil society and how it will improve on companies' current processes for this to ensure that all relevant players are in the room for similar discussions.

**Next steps**

16. As noted above, this draft Code does not at this stage synchronise with draft Online Safety Bill, which itself will be subject to modification and revision as it passes through the Parliamentary scrutiny process. Instead, the Code contains many clauses that set out our own approach to a truly systemic, risk-managed regulatory regime, built on an overarching statutory duty of care (rather than multiple duties, as is the Government's intent). As such, it is a commentary on the more limited approach the Government has adopted.

**Draft Code of Practice in respect of Hate Crime and wider legal harms**

**Risk Management and Assessment**

(1) Companies shall have carried out a suitable and sufficient assessment of the risk of harm arising from attacks on those with protected characteristics, people under 18 and vulnerable people arising from the operation of the service or any elements of it. This includes annual risk assessment reviews as well as emergency processes to address new or emerging risks. The issues to be covered in the assessment should be set down by the Regulator; any guidance as to format and evidence should be taken into account. The assessment shall be reviewed by the operator on an ongoing basis or, if there is reason to suspect that it is no longer valid; or there has been a significant change in the matters to which it relates; and where as a result of any such review changes to an assessment are required the operator shall make them. Such assessments shall be recorded and retained for a period of not less than three years or as set out by the regulator in guidance.

(2) Companies shall implement appropriate "safety by design" technical and organisational measures including but not limited to those detailed below to minimise the risks of those harms arising and mitigate the impact of those that have arisen, taking into account the nature, scope, context and purposes of the online platform services and the risks of harm arising from the use of the service;

(3) Companies shall carry out or arrange for the carrying out of such testing and examination as may be necessary for the performance of the duty of care in respect of harms arising from attacks on those with protected characteristics, bearing in mind respect for the human dignity of people involved or affected by those tests, as well as ethical considerations relating to experiments involving human participants;

(4) Companies shall regularly review and update when appropriate technical and organisational measures implemented under this code. Companies shall also keep the appropriateness and effectiveness of such measures under review during the period the online platform service is offered.

(5) Companies should ensure and be able to demonstrate their systems are safe by design, including addressing the following concerns;

(a) That algorithms do not cause foreseeable harm through promoting hateful content, for example by rewarding controversy with greater reach, causing harm both by increasing reach and engagement with a content item;

(b) That speed of transmission has been considered, for example methods to reduce the velocity of forwarding and therefore cross-platform contamination;

(c) Use of tools by actors creating or spreading harms against those with protected characteristics in order to cause harm, which are able to operate owing to weak platform polices on enforcement and moderation. This includes but is not limited to

bots, bot networks, deep fake or audio-visual manipulation materials and content embedded from other platforms;

(d) An appropriate approach to the principle of knowing your client [KYC] to address harms spread by those using false or anonymous identities;

(e) Consideration of the circumstances in which targeted advertising may be used and oversight over the characteristics by which audiences are segmented;

(f) Systems for cross-platform co-operation to ensure knowledge about repeat offenders that may present a foreseeable risk of harm in relation to attacks of those with protected characteristics;

(g) Use of tools including but not limited to prompts which clarify an individual's intended search;

(h) Policies concerning advertising sales in respect of promoting harmful content or for malicious intent in respect of those with protected characteristics.

**Enforcement**

(1) Companies should have methods to proactively identify content or activity constituting criminal hate speech, to either prevent it being made publicly available or prevent further sharing.

(2) Companies must have in place Terms and Conditions which are clear, visible and understandable by all likely users. The Terms and Conditions must also be fit for purpose for their compliance with the statutory duty of care.

(3) Companies must have reporting processes that are fit for purpose in respect of hate crime and wider harms, that are clear, visible and easy to use and age-appropriate in design. Thought should be given to reporting avenues for non-users.

(4) Companies must have in place clear, transparent and effective processes to review and respond to content reported as illegal and harmful.

(5) Companies must have in place effective and appropriate safeguards in full respect of fundamental rights, freedom of expression and relevant data protection regulation. This includes, specifically, taking reasonable steps to ensure users will not receive recommendations to criminal, hateful or inappropriate content.

(6) Companies must have in place Community Guidelines explaining their policies (and how these are developed, enforced and reviewed, plus the role of victims' groups and civil society in developing them) on harmful content, including what activity and material constitutes hateful content, including that which is a hate crime, or where not necessarily illegal, content that may directly or indirectly cause harm to others. This includes:

(a) Terrorist content;

(b) Child Sexual Exploitation and Abuse (CSEA) content;

(c) Abuse, harassment and intimidation;

(d) Stalking;

(e) Hate speech;

(f) Content promoting hostility or incitement to hatred based on legally protected characteristics whether in isolation or an intersectional manner;

(g) Disinformation where this creates the promotion of hostility or incites hatred based on legally protected characteristics;

(h) Criminal activity.

(7) Companies must have in place sufficient numbers of moderators, proportionate to the company size and growth and to the risk of harm who are able to review harmful and illegal content and who are themselves appropriately supported and safeguarded. Machine learning and Artificial Intelligence tools cannot wholly replace human review and oversight. Companies should have in place processes to ensure that where machine learning and artificial intelligence tools are used, they operate in a non-discriminatory manner and that they are designed in such a way that their decisions are explainable and auditable.

(8) Companies must have a disaggregated notification system for each type of harmful and illegal content to ensure the correct moderators, trained in their specialist subjects and on related language and cultural context considerations (where proportionately reasonable), are able to review the appropriate content, and for transparency purposes.

(9) When receiving a notification of content, a Company must review such a report taking into account national laws and the Terms of Service.

(10) A company must remove content that has been deemed to be illegal within 24 hours of becoming aware of such content. Awareness begins at the time flagged content, by means of email, in-platform notification or any other method of communication, is received.

(11) A Company must take action, proportionate to risk, on content which is not deemed to be illegal but is considered to break their Terms of Service or Community Guidelines [as soon as it is identified and no later than 24 hours]. Acceptable actions on a piece of content which violates a Company's Terms of Service can include –

(a) Removal of content;

(b) Termination of account;

(c) Suspension of account;

(d) Geo-blocking of content;

(e) Geo-blocking of account;

(f) A strike, if a strike system is in place.

(12) Without prejudice to (10), companies must have clear guidelines as to what constitutes an expedient time frame for the removal of (or temporarily limiting access to) hateful content and comply with them.

(13) Companies must have systems to prevent and identify those abusing and misusing services to create harm, including persistent abusers across a range of harms and those using anonymous accounts to abuse others.

(14) Companies must put in place systems of assessment and feedback to the initial reporter and the owner of content that has been flagged and actioned to ensure transparency of decision making. Users should be kept up to date with the progress of their reports and receive clear explanations of decisions taken.

**Outsourced Content**

(1) Companies that outsource any part of their business, including moderation of content, applications, GIFs, images, or any other content or tools, must ensure the Vendor adheres to the Terms of Service and Community Guidelines of the company and that they have employee and mental health protection policies in place that adhere to the same standard.

(2) Processes must be in place for users to report content provided by a Vendor which is illegal or violates the Company's Terms of Service or Community Guidelines.

(3) Companies must ensure adequate information is available to the Vendors on their Terms of Service and Community Guidelines to pre-empt any violations.

**Right of Appeal**

(1) Companies must put in place a Right of Appeal on all decisions made concerning illegal or harmful content, or content that has been flagged as illegal or harmful content.

(2) All users must be given a right to appeal any termination of service, suspension, geo-blocking or removal of content, whether in full or in part. Users must be able to present information to advocate their position.

(3) Companies are to ensure that Protected Speech is not removed from the platform unduly.

(4) Companies must acknowledge an appeal request, within 24 hours of receipt. If more time is needed to assess the content, the user must be informed.

(5) Appeals must take no longer than seven days to assess, except in exceptional circumstances. Exceptional circumstances could include a major disaster, or an event or incident of the same magnitude.

**User Support**

(1) Companies should provide advice and tools for "digital self-care" such that users can take steps to protect themselves in the first instance from exposure to hateful content and that these are built into new features.

(2) Companies must take steps to ensure that users who have been exposed to hateful material are directed to, and are able to access, adequate support. Support can include –

(a) Signposting and access to websites or helplines dealing with the type of hatred viewed by the user or witnessed by others who may be affected by the content, even if not the designated target;
(b) Information from, and contact details for, services providing victim support or mental health support after being exposed to hateful and harmful materials;
(c) Strategies to deal with being exposed to hateful material.

**Transparency**

(1) Companies must engage in regular self-auditing to ensure compliance with the Code of Conduct. Companies must provide quarterly transparency reports, based on these audits, of content removed which can viewed in the United Kingdom, regardless of the content's origin or origin determined by IP address. Thought should be given to effective ways to communicate with users about these reports.

(2) Transparency reports must include data and statistics produced by the IT companies. The reports must include information on –

(a) Content removal;
(b) Content removal, disaggregated by removal reason;
(c) Content removal by first source of detection. This must include both automated flagged and human flagging;
(d) Removal times, disaggregated by removal times;
(e) Appeals, disaggregated by reason for reinstatement;
(f) Content that has been reinstated;

(g) Law enforcement requests for removal;
(h) Court orders for content removal and adherence to court orders;
(i) Number of requests for removal by the Regulator;
(j) Regulator interventions.

(3) Companies must have in place the ability to grant independent scholars, academics, researchers and others with genuine, verifiable research interests that are independent of the company, the ability to access anonymous data in order to better understand the situation of illegal and harmful content, adherence to the Code of Conduct and other online activity.

(4) Companies must be able to provide information about the prevention and identification of abuse and misuse of services, including persistent abusers across a range of harms; and those using anonymous social media accounts to abuse others. This must be provided in full to government and law enforcement agencies upon request within one month of receiving the request.

**Governance and Authority**

(1) IT Companies must have in place a point of contact for law enforcement authorities. The contact is responsible for giving information about illegal content to law enforcement authorities. This includes –

(a) Information about the content;
(b) The details of the user, including location;m
(c) Details of enforcement action on the content undertaken by the IT Company;
(d) Other materials relevant to criminal investigations.

(2) Information requested by government and law enforcement authorities must be delivered within one month of receiving the request. In exceptional circumstances this can be extended, with written approval from the relevant authorities placing the request, with a full expected time frame set out.

(3) Protections must be put in place by IT Companies to ensure flagging and court orders are not used for nefarious purposes by Government agencies or law enforcement of any kind to remove content they find objectionable, which is neither illegal nor harmful.

**Education and Training**

(1) IT Companies must put in place appropriate, updated education and training for moderators, designed in consultation with independent Trusted Flaggers to insure diversity and inclusion.

(2) Materials used for training on illegal and harmful content must be made available to the Government, the Regulator, law enforcement authorities and Government agencies

upon request.

(3) IT Companies must have in place an appropriate, independent Trusted Flagger programme. The programme must include Non-Government Organisations and other experts, who will be vetted, to inform on policy development and report on new trends in harmful and illegal content. In order to ensure an effective 10 working relationship with members of Trusted Flagger programmes, IT Companies must –

(a) Ensure Trusted Flaggers are not used as a sole provider of flagging content;
(b) Ensure Trusted Flaggers are appropriately compensated and incentivised for work provided to IT Companies to ensure their compliance while not compromising their independence and impartiality;
(c) Hold regular meetings (with members of the Trusted Flagger programmes) to review content decisions and discuss any concerns;

(d) Provide support for Trusted Flaggers who are exposed to harmful content, as per the support provided to the companies own moderators, whether directly employed or working for out-sourced companies.

(4) IT Companies must provide educational tools and guidelines on their Terms of Service and Community Guidelines to ensure users are aware of permitted content on the platforms.

**Planning and Review**

(1) Companies must have plans for continually reviewing their efforts in tackling hateful material and adapt internal processes accordingly, to drive continuous improvement. This might include engagement with relevant experts or organisations to advance policy development.

(2) Companies should have intelligence systems for investigating harms organised off-platform for attack of users on a given platform and should actively and regularly share such intelligence, when received, with other platforms.

(3) Users must be given the ability to submit third-party content to the Companies' intelligence systems in relation to specific cases of content violation.

**Enforcement of this code of practice**

(1) The Regulator will assess whether actions taken by a company to comply with this code of practice are suitable and sufficient to address the risk of harm arising from hate crime in the operation of that company's service. The regulators assessment will be informed by a dialogue with a range of actors including the company concerned, victims of harm, civil society actors, the regulator's own research and any other information that the regulator considers to be relevant.

(2) Should the regulator find that a company has not taken suitable and sufficient steps under this code and harm from hate crime has occurred the regulator will employ its enforcement management model to consider whether sanctions under the online harms regime are required.

**Glossary**

**Flagged**: When a piece of content is reported to an IT company.

**Harmful Content**: Any content which is harmful to users of IT companies and wider British society. This includes activities that incite or engage in violence, intimidation, harassment, threats, defamation or other hostile act based on protected characteristics.

**Illegal Content**: Any content deemed to be illegal under British Law. This can include harmful content.

**IT Companies / Company**: Any platform accessible via the internet which allows for user-generated content to be hosted by the platform. Content can include text, imagery, photographs, videos, comments, sound and performances. Companies can include social media platforms, video sharing platforms, event scheduling/ticketing platforms, public chat services or group communications, websites, blogs or message boards.

**Moderator**: Any person who reviews content for the IT Companies. This can include third party moderators, vendors, to who the IT Companies outsource moderation and review tasks.

**Regulator:** The Government's online harms regulator.

**Trusted flagger:** an individual or organisation with particular expertise and responsibilities for the purposes of tackling illegal content online.

**Vendor**: Any external company, application, tool or other contracted out service which is available to users on the platform. This can include, but is not limited to, images, content moderation.

**Further recommended reading:**

The European Code of Conduct on Illegal Hate Speech Online
https://ec.europa.eu/info/sites/info/files/code_of_conduct_on_countering_illegal_hate_speech_online_en.pdf

Change the terms: Reducing Hate Online, Recommended Policies
https://www.changetheterms.org/terms

Digital Economy Act Code of Practice https://www.gov.uk/government/publications/code-of-practice-for-providers-of-online-social-media-platforms

Age Appropriate Design Code: https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/age-appropriate-design-a-code-of-practice-for-online-services/

AVMSD/ VSP provisions: https://www.ofcom.org.uk/tv-radio-and-on-demand/information-for-industry/vsp-regulation

The Alan Turing Institute: How much online abuse is there? A systematic review of evidence for the UK: https://www.turing.ac.uk/sites/default/files/2019-11/online_abuse_prevalence_full_24.11.2019_-_formatted_0.pdf

*17 September 2021*