DCMS SELECT COMMITTEE - CARNEGIE UK EVIDENCE RE DRAFT ONLINE SAFETY BILL

September 2021

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 the 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 an independent 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 - such as health and safety regulation - 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, risk-based and necessarily systemic. 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 (OSB) and are well aware of the complexity and the challenges with which Ministers and officials have grappled in its development. However, we are concerned that a systemic approach (evident in some parts of the Bill) has been eroded elsewhere 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 draft Bill complex and risks rendering it ineffective at tackling the root causes of harm online[4].

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

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/
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/

How has the shifting focus between 'online harms' and 'online safety' influenced the development of the new regime and draft Bill?

The Government has narrowed the scope of the online harms project from a more general approach to ensuring online safety towards a focus on a particular set of harms. So, despite being renamed as the Online Safety Bill in December 2020, the project is now more about specific harms than general safety. By seeking to specify a narrow set of harms and trying to distinguish different responses in relation to each category, the Government has made the regime more complex for the companies it proposes to regulate, the regulator who will enforce it and the users it seeks to protect, as well as opening the regime to risks of gaps and inconsistencies.

It also makes the regime less flexible. The Government now has the narrow window of the Secretary of State's Priority Harms through which to tackle new issues arising outside the defined "harms to individuals" envelope. Some of these harms are significant. For example, 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. If the Priority Harms are the route through which these are to be addressed, then we urge the Secretary of State to make an indication of "Priority Content" by the end of 2021 to assist scrutiny (although see our comments in our recent blog post, included below, on the Secretary of State's powers re reforming the "Priority Content" approach).[5]

As we have argued throughout our work, it would be less complex if the Government were to define a broader statutory duty of care (instead of three sets of more limited duties) and then within that for OFCOM to determine priorities for action, based on evidence that it presents to Parliament and the executive. (We are drafting an amendment for a general, overarching duty of care and will share this with the Committee as soon as it is available.) Reducing the emphasis on types of content (which are proxies for the harm caused) and refocussing the regime on the hazards created or amplified through the operation of the service allows for a more sophisticated range of responses that move us away from discussions about the legitimacy of take down, and move towards reducing the root causes of problems that the companies seem to have been content to ignore over the years, as recent reporting in the *Wall Street Journal* amply demonstrates.

Is it necessary to have an explicit definition and process for determining harm to children and adults in the Online Safety Bill, and what should it be?

We note that there have been some concerns about the need to specify the harms more precisely, or even that the regulation of non-criminal content is in some way illegitimate. In our view, this concern is overblown. It is a misapprehension that the draft OSB is unique in trying to protect people against content that is harmful but does not trigger the criminal law threshold. Regulation is used in many sectors as a tool for intervention that is less severe than the use of criminal sanctions; it is therefore a more proportionate response.

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

While systems regulation is not content regulation, we turn to media regulation as a good comparator because it, like social media, has an impact direct within the home (indeed the ongoing connection of people with social media via mobile devices might be said to be greater) and the content available via social media increasingly has a role in forming people's world views. It is unsurprising then, as part of wider tech exceptionalism, that technology advocates seek to demonise something which is routine in other media industries.

Current media regulation prohibits content that is harmful[6], as well as the broader category of content which 'might seriously impair the physical, mental or moral development' of those under 18, leaving the regulator (OFCOM) to give more detailed guidance as to what that means. So, while harm is not defined or quantified in the Communications Act, OFCOM's code breaks matters down into more specific instances of types of harm. For children, the Code flags, for example, the use of illegal drugs, abuse of drugs and alcohol; violence; and easily imitable dangerous behaviour. A separate section deals with hatred and abuse. Where specific concerns have arisen, OFCOM has commissioned research to produce more detailed guidance. An example is the research on 'health and wealth' claims dealing with, for example, alternative medicine claims.[7] Similarly, advertising regulation prohibits harmful content[8] and, moreover, prohibits the advertising of certain ('legal') products (eg tobacco).

It should be recognised that while some content may be prohibited, in many cases the operation of the regulatory system is not about prohibition but about giving information, or allowing views to be informed. So the BBFC provides age ratings; the watershed on linear TV operates to limit risk that content dealing with adult themes is seen by young children. Even where the Broadcast code seeks to protect views from harm, techniques OFCOM envisage include warnings to viewers, or the provision of alternative views.

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 into the impact of media upon people to carry out these duties (the health and wealth claims noted above is one example), and have revised their codes in response to changes in societal expectations. They have a long history of proportionate regulation and legal challenges (including those based on freedom of expression) have failed.

As noted above, the proposed regime should not be a content-based regime (which the current structure risks it looking like); rather it should look at how the platforms' design and operational choices influence the creation and flow of content or encourage behaviours that might (either taken item by item or collectively) be harmful. 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 those expectations in codes of practice (which may also improve the inclusiveness of the online environment). The codes of practice should focus particularly on the vectors for harm that are enabled by service design.

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). Thus the risk assessments

6 Section 319 Communications Act 2003

7 R Fincham, I Sparham, J Smith and S Grew, Health and wealth claims in programming: audience attitudes to potential harm (Optimisa Research-The Big Boutique, 2017), available: https://www.ofcom.org.uk/__data/assets/pdf_file/0016/104650/Health-claims-report.pdf

8 Rule 4 CAP Non-Broadcast Code

should not be divided artificially by content categories. At a minimum, the risk assessment should be a general risk assessment, but - as we suggested above - a general duty of care should be inserted into the draft bill.

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 and there, 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.[9]

We also note the proposals from the Law Commission regarding the communications offences. 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. The thresholds should be at least the same as those currently in force in relation to video-sharing platforms in the Communications Act.

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 the word content is the same whether referring to either a single item or to multiple items). The case of the abuse directed towards the black England footballers is a case in point. While some examples would reach the criminal threshold, it is far from clear that all would (eg instances of monkey or banana emojis), yet the cumulative impact is great - not just on the individuals receiving the communications but on society as a whole.

A similar point could be made about 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, that content and the system that caused it to be delivered might fall outside the regime. This example is one where content that in itself is not particularly harmful 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 and volume (over which the speaker has no control).

Does the draft Bill focus enough on the ways tech companies could be encouraged to consider safety and/or the risk of harm in platform design and the systems and processes that they put in place?

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[10] 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.[11]

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

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

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

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)[12] -

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.*"[13]

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[14]); navigation 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-

12 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

13 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

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

intrusive in the first place, as recognised by the UN Special Rapporteur on Freedom of Expression in the context of hate speech[15]) 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 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 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

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

What are the key omissions to the draft Bill, such as a general safety duty or powers to deal with urgent security threats, and (how) could they be practically included without compromising rights such as freedom of expression?

Harms to adults on the largest platforms are not well covered (Clause 11). The Government needs to spell out how huge volumes of - for example - racism, misogyny, antisemitism – that are not criminal but are oppressive and harmful, particularly to prominent figures – will be addressed.

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 that is not criminal bearing in mind the nature of the platform and its users[16], which seems to be the policy intent. We have two main concerns with the way this has been implemented, however.

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 the challenge of this type of language e.g. the current case with the Irish Information Commissioner of use of the term "handling" – many cases were deemed to be "handled" by not taking a decision). Contrast the position for the children's safety duty where the obligation is to "mitigate and effectively manage" risks (cl 10(2)).

Second, it is important to remember that safety duties are not just about moderation and take down. For example, a platform that wanted to adopt a more '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. It is unclear the extent to which the provisions outlining the effect of the codes (cl 37) (which should reflect the online safety objectives in clause 30) cut down platforms' choice in this context, especially taken against the context of a deficient or wilfully blind risk assessment.

This part of the Bill relies upon platforms' enforcement of their own terms of service (as against users). In so doing, it loses close connection with the characteristics of the platform design and operation and their impact on content creation (e.g. through financial or other incentives), information flows and user empowerment (e.g. through usable curation tools) that flows from a systemic approach. By contrast, the illegal content and child safety duties emphasise the importance of these "characteristics".

We also 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

16 This is a point we made in our 2019 original report, p. 42: https://d1ssu070pg2v9i.cloudfront.net/pex/pex_carnegie2021/2019/04/06084627/Online-harm-reduction-a-statutory-duty-of-care-and-regulator.pdf
R Fincham, I Sparham, J Smith and S Grew, Health and wealth claims in programming: audience attitudes to potential harm (Optimisa Research-The Big Boutique, 2017), available: https://www.ofcom.org.uk/__data/assets/pdf_file/0016/104650/Health-claims-report.pdf

disinformation flows. Civil regulation supports national security, with democratic oversight, in other regulated sectors and should do so here.[17]

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[18] 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[19]. The Cell should be put on a formal statutory footing with an obligation to report to Parliament and to include OFCOM in its work.

The proliferation of misinformation and disinformation also has a corrosive effect on the country's "epistemic security",[20] on people's ability to access and indentify 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.

Finally, fraud is another key omission. 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 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[21], Bank of England[22], City of London Police[23], Work and Pensions Committee and Treasury Committee[24], who have all commented that the scope of the Online Safety Bill should be expanded to include fraud carried out via online advertising.

17 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)

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

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

20 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/

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

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

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

24 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/

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.[25]

Are there any contested inclusions, tensions or contradictions in the draft Bill that need to be more carefully considered before the final Bill is put to Parliament?

We have reviewed the powers in the draft OSB for the Secretary of State. We set out the text of our recent blog post here.

The draft Online Safety Bill gives too many powers to the Secretary of State over too many things.[26] This is a rare point of unity between safety campaigners, who want tough legislation to address hate crime, mis/dis-information and online abuse 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[27] 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[28], 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[29], 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[30], 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[31] also do not touch content.

25 https://www.carnegieuktrust.org.uk/blog-posts/online-harms-interlocking-regulation/
26 https://www.carnegieuktrust.org.uk/annex-b-the-role-of-the-secretary-of-state/
27 https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=09000016804e0322
28 https://www.osce.org/files/f/documents/e/0/101257.pdf
29 https://www.legislation.gov.uk/ukpga/2003/21/section/319
30 https://www.legislation.gov.uk/ukpga/2017/30/contents
31 https://www.legislation.gov.uk/ukpga/2006/36/section/5

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.

(The published blog post, extracted above, is here: https://www.carnegieuktrust.org.uk/blog-posts/secretary-of-states-powers-and-the-draft-online-safety-bill/)

There is also an issue arising from the omission 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)[32].

What are the lessons that the Government should learn when directly comparing the draft Bill to existing and proposed legislation around the world?

Many democracies are now considering how to regulate technology companies, specifically social networks[33], 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[34] and OECD Guidance for Responsible Business Conduct[35] - 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; as we have already noted, 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[36]. The UK Presidency of the G7 achieved an extremely encouraging text on internet safety earlier this year,[37] which it is following through. 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.[38]

Carnegie UK has worked with a wide range of international actors on online safety – below, we draw out some further lessons learned from that for the United Kingdom in how it approaches international issues in regulation. This is based upon our evidence[39] to the ongoing

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

33 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

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

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

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

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

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

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

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 General Comment 25 on the Convention on Rights of the Child concerning children's rights in relation to the digital environment[40].
- 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.[41]
- 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 was the G7 tech ministers' declaration[42], secured by Oliver Dowden; and the next the recent G7 Interior and Security Ministers declaration[43]. 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[44] and Rule of Law programmes.
- Improve the byzantine, even chaotic UN process (WSIS,[45] 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.

Carnegie UK
September 2021
contact: maeve.walsh@carnegieuk.org

40 https://www.ohchr.org/EN/HRBodies/CRC/Pages/GCChildrensRightsRelationDigitalEnvironment.aspx
41 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
42 https://www.gov.uk/government/publications/g7-digital-and-technology-ministerial-declaration
43 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
44 https://thecommonwealth.org/commonwealth-cyber-declaration
45 https://www.itu.int/net/wsis/