TACKLING ONLINE ABUSE: WRITTEN EVIDENCE SUBMITTED BY CARNEGIE UK TRUST ON 13/10/2021 (TOA0016)

In 2018, the Petitions Committee held an inquiry into online abuse and the experiences of disabled people. Despite overwhelming evidence of the scale of the issue, for disabled people and for many other people on the internet, there has been little progress in making the internet a safer place to be. Petitioners continue to share their experiences of abuse online, most recently The Only Way Is Essex star Bobby Norris who petitioned to make online homophobia a specific crime and for new technological solutions to be used to take online trolls off social media. The Petitions Committee is launching a new inquiry into tackling online abuse, to examine what progress has been made since our predecessor Committee's work on the issue and to press the Government on the action it needs to take.

The Committee will consider:

• The scale and impact of online abuse on internet users, including disabled people, the LGBT+ community and other minority groups

• Government proposals to tackle the issue, including the Online Harms White Paper

• Legal and technological solutions to take action against people who commit online abuse

1. In the light of the resumption of the Petition Committee's inquiry into Tackling Online Abuse, we welcome the opportunity to submit revised evidence. This submission sets out the background to our work on the development of a statutory duty of care for online harm reduction and responds primarily to the latter two considerations set out in the inquiry's scope, taking into account the fact that the Government has since updated its proposals since the publication of the Online Harms White Paper, with its draft Online Safety Bill currently undergoing pre-legislative scrutiny in Parliament.

2.

Background

3. Carnegie UK Trust (CUKT) is a not-for-profit organisation focused on improving wellbeing through a range of research, advocacy and community programmes. Since early 2018, it has supported work on new proposals for internet harm reduction instigated by William Perrin (a former UK Civil Servant, who is now a Carnegie UK Trustee) and Professor Lorna Woods (Professor of Internet Law, University of Essex, and an EU national expert on free speech and communications regulation). Their work has focussed on the development of a statutory duty of care to reduce reasonably foreseeable harms on social media enforced by a regulator.

4. Our proposals have been published in a series of blogs and publications for Carnegie and developed further in evidence to Parliamentary Committees[1]. The Lords Communications Committee[2] and the Commons Science and Technology Committee[3] both endorsed the Carnegie model, as have a number of civil society organisations[4]. In April 2019, the

[1] Our work, including blogs, papers and submissions to Parliamentary Committees and consultations, can be found here: https://www.carnegieuktrust.org.uk/project/harm-reduction-in-social-media/

[2] https://www.parliament.uk/business/committees/committees-a-z/lords-select/communications-committee/inquiries/parliament-2017/the-internet-to-regulate-or-not-to-regulate/

[3] https://publications.parliament.uk/pa/cm201719/cmselect/cmsctech/822/82202.htm

[4] For example, NSPCC: https://www.nspcc.org.uk/globalassets/ documents/news/taming-the-wild-west-web-regulate-social-networks.pdf; Children's Commissioner: https://www.childrenscommissioner.gov.uk/2019/02/06/childrens-commissioner-publishes-astatutory-duty-of-

government's Online Harms White Paper[5], produced under the then Secretary of State for Digital, Culture, Media and Sport, Jeremy Wright, proposed a statutory duty of care enforced by a regulator in a variant of the Carnegie model. France[6], and apparently the European Commission, are now considering duty of care models for online harms (the proposal for a Digital Services Act refers to due diligence obligations).

5. In December 2019, while waiting for the Government to bring forward its own legislative plans, we published a draft bill[7] to implement a statutory duty of care regime, based upon our full policy document of the previous April[8]. Our recent publications, with specific relevance to this inquiry, include: the draft Hate Crime Code of practice[9], which we developed with a number of civil society organisations, including Antisemitism Policy Trust; our blog post on the racist abuse of England footballers[10]; our initial analysis of the draft Online Safety Bill[11]; and our evidence to both the Joint Committee on the Online Safety Bill[12] and the DCMS Select Committee inquiry into online harms and online safety[13].

Government proposals to tackle online harms

6. Government action on online harms is long overdue. While we welcome the publication of the draft Online Safety Bill and the focus that it is now receiving via multiple Parliamentary Committees, it has taken four years for the Government to get to this point – and will likely take at least another two before the regime is in force and Ofcom can start to regulate.

7. We regret the delays to the Government proposals, particularly given the evidence of an upsurge during lockdown in many of the harms that the Bill would cover. For example, child sexual abuse and exploitation (which are criminal offences) have been exacerbated by the Covid19 crisis as more and more time is spent online at home and children and young people's unsupervised social media activity increases. In her evidence to the Home Affairs Committee, the Home Office Minister, Baroness Williams, referred to a "21% uptick" in hate crime during the lockdown period – and the recent abuse of black England footballers after the European Championship Final has exposed the limitations of social media platforms' existing processes in dealing with such activity. The former Digital Minister, Caroline Dinenage, also referred in Parliament to evidence of a rise during the pandemic in incidents of revenge porn and sexploitation.[14]

---

care-for-online-service-providers/; Royal Society for Public Health: https://www.rsph.org.uk/our-work/policy/wellbeing/new-filters.html

[5] https://www.gov.uk/government/consultations/online-harms-white-paper

[6] http://www.iicom.org/images/iic/themes/news/Reports/French-social-media-framework---May-2019.pdf

[7] https://www.carnegieuktrust.org.uk/publications/draft-online-harm-bill/

[8] https://d1ssu070pg2v9i.cloudfront.net/pex/carnegie_uk_trust/2019/04/08091652/Online-harm-reduction-a-statutory-duty-of-care-and-regulator.pdf

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

[10] https://www.carnegieuktrust.org.uk/blog-posts/racist-abuse-of-footballers-using-social-media-and-the-draft-online-safety-bill/

[11] https://www.carnegieuktrust.org.uk/blog-posts/the-draft-online-safety-bill-carnegie-uk-trust-initial-analysis/

[12] https://committees.parliament.uk/writtenevidence/39242/html/

[13] https://committees.parliament.uk/writtenevidence/39491/html/

[14] We discuss below some of the difficulties around definitions of harms, particularly in relation to abusive and offensive behaviour online. Many of these are in scope of the Law Commission's ongoing review: https://www.lawcom.gov.uk/law-commission-to-undertake-phase-2-of-the-abusive-and-offensive-online-communications-project/

8. Our evidence to the Home Affairs Committee last year focused on the increasing prevalence of other harms – such as fraud and scams, and misinformation/disinformation, which (despite the recent concession on user-generated scams) are both currently outwith the Government's proposals – during the Covid 19 pandemic[15]; and our evidence to the Joint Committee explores these gaps further.

9. The Petitions Committee is right to focus at this time on the impact of online abuse – which may well fall into what the draft Bill terms 'harmful but legal content - and the means by which it can tackled. We hope that your inquiry will dovetail effectively with the pre-legislative scrutiny process so that any recommendations can be fully considered not just by that Committee but also by the Government, in its response and amended Final Bill. As has been seen already in the oral evidence to the Joint Committee (in particular that given at its first hearing by Rio Ferdinand, Kick it Out and the FA; and by Danny Stone of the Antisemitism Policy Trust[16]), the effectiveness of the government's legislative proposals to deal with online abuse are debatable, despite the undeniable impact that it has on individuals targeted by trolls, subjected to pile-ons or repeatedly harassed or intimidated by abusers who they may or may not know. The Petitions Committee will doubtless receive ample testimony from victims during the course of this inquiry. Despite this, in her evidence to the Home Affairs Committee last year, the former Minister for Digital and Culture frequently referred to the "difficult balance to strike" in dealing with such harmful or damaging information content "within the realms of people's freedom of speech". On this point, we would recommend Hope Not Hate's recent paper "Free Speech for All", which sets out the breadth of online harms and types of abuse that would be excluded if "legal but harmful" were to be dropped from the Bill.[17]

10. The Government will be under increasing pressure on this point from the tech companies and free speech advocates. We would therefore urge policymakers – and Parliamentarians – to remain focused on how a systemic duty of care, enforced by an independent regulator[18], would work in practice. As William Perrin set out in his oral evidence to the Joint Committee, the introduction of a general, overarching duty of care would help greatly in this regard and we are working on amendments to the draft Bill at present, which we will be happy to share with the Committee in due course.[19] We set out more on the necessity for a "systemic" approach to regulation below and recommend to the Committee the comprehensive paper by Professor Lorna Woods on how a statutory duty of care would fit within a framework of fundamental rights, especially freedom of expression.[20]

<u>Legal and technological solutions to take action against people who commit online abuse</u>

11. It is not only the expression of the harmful content in itself that causes problems but the design of platforms that exacerbate its spread and, in the case of online abuse of public figures or minority groups, encourage others to share it further or participate in, and amplify it,

---

[15] https://committees.parliament.uk/writtenevidence/5389/html/
[16] https://committees.parliament.uk/oralevidence/2694/html/
[17] https://www.hopenothate.org.uk/wp-content/uploads/2021/08/Free-Speech-For-All-2021-08-FINAL.pdf
[18] On this point (Ofcom's independence) we have argued in a recent blog post that the draft Online Safety Bill grants too many powers to the Secretary of State; we will be publishing amendments to the Bill to address this in due course. https://www.carnegieuktrust.org.uk/blog-posts/secretary-of-states-powers-and-the-draft-online-safety-bill/
[19] https://committees.parliament.uk/oralevidence/2796/html/
[20] https://d1ssu070pg2v9i.cloudfront.net/pex/carnegie_uk_trust/2019/12/10111353/The-Carnegie-Statutory-Duty-of-Care-and-Fundamental-Freedoms.pdf

themselves. The speed and scale of its spread and promotion – a spread encouraged and facilitated by the platforms' own system design – is what sets the nature of online abuse apart from its offline equivalent. This includes for example, their algorithms, recommender models, reliance on user profiling and micro-targeting[21], or nudges to users to like or share content without time for reflection. Indeed, the impact of Twitter's design choices was exposed starkly when, during the period earlier this year when Wiley was posting antisemitic tweets, its promoted content in its "trends" list on "Jews".[22] A significant part of the problem, in our view, relates to these information flows, and this is an aspect that does not readily fit a framework – as is the case with the draft Online Safety Bill – that is designed around a distinction on whether content is contrary to the criminal or legal but harmful.

12. We recommend that the Committee explores the impact should the Government limit the online harms regime – as was argued recently by the Lords Communications and Digital Committee – to those that lie only within the boundaries of criminal law. There are many areas where a regulatory system penalises people for things that are not criminal offences and where a regulator and the companies that are regulated are trusted to make a judgement.

13. In the UK, different types of media have been regulated for harm in different ways for decades. Current media regulation prohibits content that is harmful[23], 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 approach to COVID19 issues.[24] Similarly, advertising regulation prohibits harmful content (understood quite broadly) 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.

14. 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. 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[25]. 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.

---

[21] See the r020 report by the Centre for Data Ethics and Innovation, which recommended that online targeting be subject to the duty of care: https://www.gov.uk/government/publications/cdei-review-of-online-targeting

[22] https://twitter.com/CCDHate/status/1287055487410868225?s=20

[23] S 319 Communications Act 2003

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

[25] 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?

15. 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.

16. 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.

17. As we set out in our blog on the abuse of black England footballers, there are also more pressing actions that the Government could take to address this issue. Racism – or any other form of abuse that falls short of the criminal threshold – is weakly addressed in the Bill. There are two aspects, either of which would fall under the weak clause 11 'duties to protect adults' online safety':

    1) Such abuse can be classified by the Secretary of State as 'priority content that is harmful to adults' (and approved by Parliament). By contrast to the position with regard to the specification of priority illegal content, the Secretary of State has to consult OFCOM before making a regulation about priority content that is harmful to adults (Clause 47), but the Secretary of State is not bound to follow OFCOM's views.

       The Secretary of State has been reluctant to even give an indicative view on what priority content would be. We understand much work is underway in government to evidence priority content but have no inkling of when the Secretary of State might set some thoughts out. There is little to stop them at this early stage indicating, for example in a speech to Parliament, some of the areas that are of particular importance. It would make scrutiny much easier.

    2) The 'adults risk assessment' could also reveal racism content that is harmful to adults

In either of these cases, the weak clause 11 merely requires a company to specify in its terms of service how priority content that is harmful to adults is to be 'dealt with' by the service and that the terms of service be applied 'consistently'. It only applies to Category 1 platforms, which are likely to be just the largest platforms.

18. There are several issues with the way harm to adults is handled in the Bill.

    - **'Dealt with' (unqualified) in Clause 11 is vague, you can 'deal with' something, even if it has come up in a risk assessment, by looking at it and deciding not to do anything.** We think that here the obligation to deal with is intended to mean that if the terms of service are inadequate or are not enforced such that harm to adults occurs, then OFCOM can step in and request changes. But this is not drafted clearly and is in stark contrast to, say, the illegal content risk management clauses. We expect this to be improved or elaborated upon in the next draft of the Bill.

    - A second issue is that **relying on terms of service** suggests that the regulatory focus is at the point when companies tell users what they can and cannot do – content moderation policies and take down rules. What it **does not seem to do is to require companies to change their upstream systems and processes**, which are more likely to be effective at scale than tighter terms of service. Such mechanisms include giving

people tools to protect themselves, not algorithmically promoting racism, not recommending groups etc.

- Relying on *ex post* reactions such as take downs not only ignores the range of possible interventions (most of which would be more proportionate from a freedom of expression perspective) but gives rise to other difficulties. The speed with which the abuse of the footballers mounted up in the aftermath of the Final demonstrates how, if thousands of pieces of individual content simultaneously breach their terms and conditions, platforms using the inadequate systems they currently run will be unable to respond at the scale required to neutralise the aggregate impact. Moreover, platforms may assess the impact of each of those posts individually and fail to identify the cumulative impact of the abuse.

19. Again, we return to the fact that a more effective solution would be to have a general duty to take reasonable steps to prevent reasonably foreseeable harms and for the Secretary of State and Parliament to give an early steer as to their priorities.

20. This Government's proposed approach will not be sufficient to reduce the potential harm already experienced by many users online; nor will it prevent the emergence of harm from whatever the next global, societal or democratic crisis may be. The Carnegie April 2019 policy document[26] 'Online harm reduction – a statutory duty of care and regulator' discuss the arguments for a systemic approach at length, building on a "precautionary principle" that places responsibility for the management and mitigation of the risk of harm - harms which they have had a role in creating or exacerbating - on the tech companies themselves. In summary:

> "*At the heart of the new regime would be a 'duty of care' set out by Parliament in statute. This statutory duty of care would require most companies that provide social media or online messaging services used in the UK to protect people in the UK from reasonably foreseeable harms that might arise from use of those services. This approach is risk-based and outcomes-focused. A regulator would have sufficient powers to ensure that companies delivered on their statutory duty of care. …*
>
> *"Everything that happens on a social media or messaging service is a result of corporate decisions: about the terms of service, the software deployed, and the resources put into enforcing the terms of service and maintaining the software. These design choices are not neutral: they may encourage or discourage certain behaviours by the users of the service … A statutory duty of care is simple, broadly based and largely future proof. For instance, the duties of care in the Health and Safety at Work Act 1974 still work well today, enforced and with their application kept up to date by a competent regulator.*
>
> *"A statutory duty of care focuses on the objective – harm reduction – and leaves the detail of the means to those best placed to come up with solutions in context: the companies who are subject to the duty of care. A statutory duty of care returns the cost of harms to those responsible for them, an application of the micro-economically efficient 'polluter pays' principle … The continual evolution of online services, where software is updated almost continuously makes traditional evidence gathering such as long-term randomised control trials problematic. New services adopted rapidly that potentially cause harm to illustrate*

[26] See https://d1ssu070pg2v9i.cloudfront.net/pex/carnegie_uk_trust/2019/04/08091652/Online-harm-reduction-a-statutory-duty-of-care-and-regulator.pdf

long standing tensions between science and public policy. For decades scientists and politicians have wrestled with commercial actions for which there is emergent evidence of harms: genetically modified foods, human fertilisation and embryology, mammalian cloning, nanotechnologies, mobile phone electromagnetic radiation, pesticides, bovine spongiform encephalopathy. In 2002, risk management specialists reached a balanced definition of the precautionary principle that allows economic development to proceed at risk in areas where there is emergent evidence of harms, but scientific certainty is lacking within the time frame for decision making.[27]

"Emergent evidence of harm caused by online services poses many questions: whether bullying of children is widespread or whether such behaviour harms the victim; whether rape and death threats to women in public life has any real impact on them, or society; or whether the use of devices with screens in itself causes problems. The precautionary principle provides the basis for policymaking in this field, where evidence of harm may be evident, but not conclusive of causation. Companies should embrace the precautionary principle as it protects them from requirements to ban particular types of content or speakers by politicians who may over-react in the face of moral panic. Parliament should guide the regulator with a non-exclusive list of harms for it to focus upon. Parliament has created regulators before that have had few problems in arbitrating complex social issues; these harms should not be beyond the capacity of a competent and independent regulator. Some companies would welcome the guidance. [28]

21. Both the Twitter/Wiley case and the racist online abuse targeted at the black England footballers exemplify the case for a *systemic* duty of care for online harm reduction that makes social media platforms accountable for the effective functioning and design of their services, from carrying out robust risk assessments to identify the risk of reasonably foreseeable harm to users, to having effective and swift response and resolution systems when harms (such as the propagation of hate speech or targeted online abuse directed towards individuals or minority groups) arise. Where these systems fail, the regulator would have powers to take action.
22. We are happy to talk further to the Committee or to provide further evidence to this inquiry.

---

27 https://webarchive.nationalarchives.gov.uk/20190701152341/https://www.hse.gov.uk/aboutus/meetings/committees/ilgra/pppa.htm

28 https://d1ssu070pg2v9i.cloudfront.net/pex/carnegie_uk_trust/2019/07/04163920/Online-Harm-White-paper-.pdf