

Revise and Resubmit? Reviewing the 2019 Online Harms White Paper

Prof. Victoria Nash, Associate Professor and Senior Policy Fellow

Oxford Internet Institute, University of Oxford

Victoria.nash@oii.ox.ac.uk @VickiNashOII

As the 2019 Online Harms White Paper (OHWP) notes, the Internet is an increasingly integral part of our lives, and can offer ‘significant benefits’¹. In order to ensure these benefits are not undermined, the OHWP argues that new regulation is needed to reduce a wide array of ‘online harms’ such as those described in the joint ministerial foreword: ‘In the wrong hands the Internet can be used to spread terrorist and other illegal or harmful content, undermine civil discourse, and abuse or bully other people.’² This opening statement (and, indeed, its joint authorship³) is key to understanding many of the challenges of this policy proposal, insofar as at its heart lies an unhelpful elision between illegal and legal-but-harmful content. By conjoining these two issues the OHWP weakens its own case for new regulation, but this is by no means the only flaw of the proposed approach. Drawing on scholarship from Internet (social science) research and policy rather than media law, I highlight below the main limitations and outline suggestions for a more coherent approach.

On social ills and technological cures

The following criticisms of the current draft of the OHWP should not be read as denying a role for greater policy attention to the challenges and opportunities of the digital economy. Given the range and extent of problematic content and behaviour described in the OHWP, government concern and scrutiny is certainly merited. The problems outlined range from illegal content such as child sexual exploitation and abuse (CSEA) material, hate speech and terrorist content, to harmful but legal content such as material uploaded by prisoners, bullying, self-harm imagery and online disinformation as well as non-content related issues such as screen time and ‘designed addiction’. However, the sheer breadth of this list is immediately problematic. The idea that a single effective and proportionate regulatory approach could be designed in such a way as to tackle every one of these matters is highly presumptuous and neglects the wide array of complex social factors underpinning the production, sharing and engagement of such content. Efforts to reduce the prevalence of hate speech or CSEA, for example, could indeed benefit from a common focus on the responsibilities of online platforms to ensure that they do not host illegal material. But eradicating online CSEA would entail also addressing the demand for CSEA and the illegal economy that underpins much of its production, whilst we have a poor understanding of measures that can tackle the social roots of hate speech at all. Ultimately all that unites these ‘online harms’ is the fact that they are aspects of our online experience. To be clear then, what we have in the OHWP is a set of

¹ HM Government. (2019). Online Harms White Paper. <www.gov.uk/government/consultations/online-harms-white-paper> accessed 28 August 2019

² *ibid* 3.

³ The OHWP was published jointly by the Department for Digital, Culture, Media and Sport and the Home Office and thus reflects policy priorities of both Departments.

regulatory proposals that primarily target the *technological* manifestation of *social* ills rather than the ills themselves. Is this justified?

It is true that Internet technologies have certain features or affordances which make online content challenging to control or remove, whilst other features may serve to enhance the impact of material posted. Social media platforms provide new unmediated spaces for individuals to create and share content without prior oversight from any editorial eye⁴. Many such platforms allow users to do so whilst preserving their anonymity, which is excellent for enabling free speech but less helpful in maintaining civility⁵. ‘Online disinhibition effects’ are associated not just with anonymity but with other aspects of Internet use such as the asynchronous, geographically dispersed nature of online communication which can also diminish social cues and enable antisocial behaviours⁶. In other networks where friends or acquaintances exchange content, there is evidence that effects of ‘social contagion’ make us more likely to share or believe ‘trusted’ information from such peers⁷. These factors together with other technical affordances enable online content to circulate at speed, at scale and without any points of centralised control which could enable effective removal of illegal material. To this extent then, a focus on improving platform governance could be justified if it demonstrated both a clear policy rationale and a carefully-calibrated set of proposals. Arguably, neither of these is evident in the OHWP.

Justifying new regulatory measures

In terms of policy rationale, there are effectively three main arguments set out in the Introduction to the OHWP, relating to illegal and harmful content respectively⁸. They are that:

- The continued prevalence of serious illegal content and activity online is unacceptable;
- Harmful content and activity is damaging for individuals, particularly for children and young people;
- There is increasing public concern about online harms.

I won’t address the third of these claims, as even OfCom’s own data shows that far more Internet users are concerned about online data theft and fraud than they are about cyber-bullying or CSEA content⁹.

Illegal content

It is hard to disagree with the first of these arguments, insofar as the quantity of CSEA material or hate speech available online in the UK as detected by the Internet Watch Foundation, makes clear¹⁰.

⁴ Nicole B Ellison & danah boyd, ‘Sociality through Social Network Sites’ in William H Dutton (ed), *The Oxford Handbook of Internet Studies* (Oxford University Press 2013).

⁵ Zizi Papacharissi, ‘Democracy Online: Civility, Politeness, and the Democratic Potential of Online Political Discussion Groups’ (2005) 6 *New Media & Society* 259.

⁶ John Suler, ‘The Online Disinhibition Effect’ (2004) 7 *CyberPsychology and Behaviour* 321.

⁷ Robert M Bond and others, ‘A 61-million-person Experiment in Social Influence and Political Mobilisation’ (2012) 489 *Nature* 295.

⁸ OHWP 11.

⁹ OfCom, *Online Nation* (30 May 2019) < www.ofcom.org.uk/data/assets/pdf_file/0028/149068/online-harms-chart-pack.pdf > accessed 28 August 2019. This observation was noted in Victoria Baines, *On Online Harms and Folk Devils: Careful Now* (24 June 2019) < medium.com/@vicbaines/on-online-harms-and-folk-devils-careful-now-f8b63ee25584 > accessed 22 August 2019.

However, it is challenging to prove the consequent claim that ‘existing efforts to tackle this activity have not delivered the necessary improvements’¹¹ or that the problem is worsening, on the basis that we are finding more illegal material online. As we develop ever more sophisticated tools for automatically identifying CSEA material online¹², we may just be finding more of it, rather than this reflecting an increase in scale of the source material. This would be a sign that at least some existing efforts are working. For example, Facebook reportedly found a staggering 8.7 million images of CSEA on its platforms in 3 months in 2018¹³. But how are we to interpret this? As evidence that the current regulatory framework isn’t working (because so much was found on the platform), or that it is working (because the platform took measures to identify the content and remove it)? Certainly, Facebook chooses to disclose only a limited amount of information about its processes, which makes it hard to answer such a question. And if we think that efforts by users to upload or share such quantities of CSEA material indicate the current regulatory system isn’t working, is this because of a failure by online platforms to fulfil their legal responsibilities, because there is a gap in the existing regulatory framework governing those platforms or because of other failures, such as failure to tackle the underlying supply of and demand for CSEA content? The answers to these questions matter insofar as they affect not only the rationale for policy intervention but also the goal and nature of interventions.

Of course, illegal content such as CSEA has no legitimate place on our online platforms, and as such, renewed policy efforts to tackle the circulation of CSEA, hate speech and terrorist material online are welcome. However, there is a tension in the OHWP between a well-justified determination (I presume led by the Home Office) to tackle such illegal activity, and a more experimental approach to regulating online companies (I presume led by the DCMS). As a result, neither aim is well met. To the extent that the OHWP treats these CSEA, hate speech or terrorist material as just illegal content on platforms rather than illegal behaviour or activity by individuals, there is a risk that any new measures will tackle only partial aspects of the underlying problem. In an ideal scenario, serious new measures to tackle illegal online content would be presented as just one aspect of a wider ‘public health’ approach seeking to alter the whole environment in which hate crimes or child sexual exploitation arise, rather than just focusing on platform responsibilities¹⁴. Such a systemic approach would justify increased use of technological means to disrupt online creation and sharing but would situate this alongside population-wide behaviour and norm changes to reduce overall levels of risk whilst revisiting effective legal frameworks which discourage and punish offending. Vitally, the online component of any new measures should take into account the particular social, economic, political and technological drivers which underpin the online manifestation of these crimes – they are so much more than just content problems.

¹⁰ Internet Watch Foundation, Annual Report (2019) <www.iwf.org.uk/report/2018-annual-report> accessed 26 July 2019.

¹¹ OHWP 12.

¹² For example, Canada’s Project Arachnid proactively crawls the web looking for CSEA content for report and removal: Cybertip.ca (2016) <cprojectarachnid.ca/en/> accessed 28 August 2019.

¹³ ‘Facebook Secret Software Reveals 8.7m Child Abuse Images on its Platform’ *The Guardian* (25 October 2018) <www.theguardian.com/technology/2018/oct/25/facebook-8m-child-abuse-images> accessed 26 July 2019.

¹⁴ Megan Clarke and others, ‘A Public Health Approach to Addressing Internet Child Sexual Exploitation’ in Ethel Quayle & Kurt M. Ribisl (eds) *Understanding and Preventing Online Sexual Exploitation of Children*, (Routledge 2012).

Legal but harmful content

If Home Office efforts to tackle serious online crimes are thwarted by the OHWP's lack of focus and depth, the broader aims of DCMS to explore new ways of regulating online companies is muddled by the focus on harm. I have argued elsewhere that a policy focus on *harm* is welcome insofar as this provides a better justification for regulatory intervention than online *risk* or even exposure to risk (the 'risk of risk')¹⁵. But such a focus implies that we should have evidence of the harms caused. The OHWP provides very little evidence of the prevalence, extent or seriousness of such harms, and as such this severely weakens the case for increased intervention. Statistics are provided, we presume with the intent of demonstrating the extent of harms experienced. We learn for example, that in 2017, 'one in five children aged 11-19 reported having experienced cyberbullying in the past year'¹⁶; that 8.2% of young adults in a survey reported actively searching for information about self-harm¹⁷ and that 47% of parents are concerned about the amount of time their children spend online¹⁸. These figures are helpful illustrations of government's reasons for concern, but no effort is made to contextualise the figures or help us understand the severity of the harms experienced and how closely these are associated with online content and activity. Given that the intended outcome of the consultation is a framework of measures for regulating online content (albeit enacted at arms-length by private companies) there is a particular onus on government to demonstrate that the postulated harms are sufficient to justify the limitations that will be placed on freedom of expression and participation online.

The reliance upon assumptions of harm is further problematic because in several cases, it is unclear whether a persuasive evidence base could ever be attained. For instance, evidence for harms around eating disorder or self-harm content is very mixed¹⁹. There is evidence that this type of content may indeed be harmful for vulnerable individuals in some contexts, but not necessarily for the general population. Further there is also evidence that prohibiting the posting of such content may cut off one route that enables individuals to counter isolation, reduce self-harm urges or even find support towards recovery²⁰. Even if we agree that such content is inherently risky, the fact remains that the harm to a particular individual may need to be weighed up against its benefits to another. Whilst there is an expanding academic evidence base that helps to shed light on the risks and harms associated with different types of online content or activity, it is clear from published meta-analyses and systematic reviews of this literature that few simple causal claims can be made²¹. As such, description of examples of 'harmful content' may help us understand why government feels

¹⁵ Vera Slavtcheva-Petkova, Victoria Nash and Monica Bulger 'Evidence on the Extent of Harms Experienced by Children as a Result of Online Risks: Implications for Policy and Research' (2014) 18 *Information, Communication & Society* 48.

¹⁶ OHWP 16.

¹⁷ *ibid* 19.

¹⁸ *ibid* 21.

¹⁹ Helen Sharpe and others 'Pro-eating Disorder Websites: Facts, Fictions and Fixes' (2011) 10 *Journal of Public Mental Health* 34; Kate Daine and others, 'The Power of the Web: A Systematic Review of Studies of the Influence of the Internet on Self-Harm and Suicide in Young People' (2013) 8 *PLoS ONE* 8.

²⁰ Stephen P. Lewis and Yukari Seko, 'A Double-edged Sword: A Review of Benefits and Risks of Online Non-Suicidal Self-injury Activities' (2016) 72 *Journal of Clinical Psychology* 249.

²¹ *ibid* and also for example, Sonia Livingstone and Peter K. Smith, 'Annual Research Review: Harms Experienced by Child Users of Online and Mobile Technologies: the Nature, Prevalence and Management of Sexual and Aggressive Risks in the Digital Age' (2014) 55 *Journal of Child Psychology and Psychiatry* 635; Robin M. Kowalski and others, 'Bullying in the Digital Age: A Critical and Meta-analysis of Cyberbullying Research among Youth' (2014) 140 *Psychological Bulletin* 1073.

intervention is needed, but in the absence of a mature and rigorous evidence base, it is hard to see how this justifies the broad ambitions to regulate set out in the OHWP.

A carefully calibrated set of proposals?

If we set aside the focus on harms or illegal content, the OHWP could instead be read as a manifesto for government-led platform governance. Interpreted in this light, we should consider the normative appeal of the governance model set out, rather than just the strength of the grounds for intervention. This is a topic currently receiving much scholarly attention, addressing aspects such as the complexity of current governance arrangements²² and their adequacy in meeting the range of public policy challenges²³; the legitimacy of current modes of platform governance²⁴ and alignment with human rights²⁵.

The OHWP proposes a governance framework for relevant companies ‘that allow users to share or discover user-generated content or interact with each other online’²⁶. The question of scope arises here, as this OHWP definition potentially includes a range of companies that would not normally be described as platforms, such as online variants of traditional news media, or messaging apps, driven we assume by the focus on places where ‘online harms’ might arise. Despite this confusing (and problematic) breadth of scope, it is still possible to interpret DCMS’s aims as proposals for platform governance insofar as the model of regulation proposed seems driven by a determination to shape the practices and policies of online platforms. It also remains to be seen whether this expansive scope will ultimately be retained, as it raises several concerns relating to issues such as the freedom of the press²⁷, rights to privacy and effects on small and medium-sized businesses²⁸.

Whether or not the final scope will change, the normative foundation of the OHWP is that relevant companies will in future have a statutory ‘duty of care’. There has already been extensive commentary on the suitability of the ‘duty of care’ framework for application in the online context. As others have noted²⁹, in the *offline* context, the duty of care applies to the provider of resources such as a playground or workplace, who must ensure that these resources are not dangerous for users and may otherwise be found negligent. The provider of those resources is usually proximate to

²² Natali Helberger, Jo Pierson and Thomas Poell, ‘Governing Online Platforms: From Contested to Cooperative Responsibility’ (2018) 34 *The Information Society* 1; Robert Gorwa, ‘What is Platform Governance?’ (2019) 22 *Information, Communication & Society* 854.

²³ Victoria Nash and others, ‘Public Policy in the Platform Society’ (2017) 9 *Policy & Internet* 368.

²⁴ Nicolas Suzor, Tess Van Geelen & Sarah Myers West, ‘Evaluating the Legitimacy of Platform Governance: A Review of Research and a Shared Research Agenda’ (2018) 80 *International Communication Gazette* 385.

²⁵ *Ibid*; Rikke F. Jørgensen, ‘What Platforms Mean When They Talk About Human Rights’, (2017) 9 *Policy and Internet* 280.

²⁶ OHWP 8.

²⁷ Matthew Moore, ‘Fears for Press Freedom under Internet Abuse Law’, *The Times* (2 July 2019) <www.thetimes.co.uk/article/fears-for-press-freedom-under-internet-abuse-law-xr65cl0zh> accessed on 28 August 2019.

²⁸ Harry de Quetteville and Matthew Field, ‘Could Tough New Rules to Regulate Big Tech Backfire?’, *The Telegraph* (9 April 2019) <www.telegraph.co.uk/technology/2019/04/09/could-tough-new-rules-regulate-big-tech-backfire/> accessed on 28 August 2019.

²⁹ Graham Smith, ‘Take Care with the Social Media Duty of Care’ (19 October 2018) <<https://www.cyberleagle.com/2018/10/take-care-with-that-social-media-duty.html>> accessed 28 August 2019.

the users and should have a clear understanding of the ways in which their resources will be used and by whom. The harms in focus are primarily safety-related harms.

In the context of *online* platforms, ‘harms’ may result when one user consumes content created or shared by another user. Apart from specific cases such as child grooming, incitement to violence or sharing of terrorist content, it is not clear that the harms in focus are predominantly safety-related. Others relate to offense, emotional distress or other less immediate psychological harms, where there is likely to be substantial variation amongst populations as to how or indeed, whether such harms will be experienced, at least when compared to risks to physical safety or health. Crucially, platforms do not create the ‘harm’, other users do, yet the duty of care is to be borne by the platform and not users. Given the current absence of a strong research evidence base, it is also not clear how platforms and other online companies could be expected to have a clear understanding of the ways in which third party content will affect other users. Add to this the further complication that users of online content may present different risk profiles for diverse sorts of content, and it becomes very challenging to imagine how platforms could exert their duty of care for all users, or indeed what would clearly count as negligence.

If viewed from a platform governance perspective, the most significant failing of the duty of care approach is that (at least as currently framed) it doesn’t tackle the unique regulatory challenges posed by problematic behaviour and content on platforms. It fails on two counts. First, imposing a duty of care fails to engage with the problem of user responsibility and second, it fails to engage systematically with the complex array of procedures which platforms use to govern online content and behaviour.

The first failing is short-sighted rather than terminal, and matters more for if the aim is reducing harms rather than providing a new framework for platform governance. In choosing to target the technological manifestation of social ills rather than the underlying behaviours themselves, the duty of care framework proposes no new sanctions for users who create or share illegal and ‘harmful’ content, (although it would operate alongside existing sanctions in private or criminal law). The OHWP does include in its exemplar codes of practice several measures that target individual user behaviour, but these are only instrumentally justified, presented as problems caused by users of particular platforms rather than a wider societal matter with potentially complex social, economic and political causes. Similarly, the proposed expansion of media literacy education is also very welcome, but again feels tacked on, rather than a crucial plank of a ‘whole-system’ approach to tackling problematic online behaviour. Such a system-wide approach would undoubtedly be more costly and more complex, but societal problems rarely have simple technological fixes.

As for the second flaw, I have elsewhere argued that the optimal approach to platform governance would involve requiring greater ‘procedural accountability’³⁰. This is defined as the ‘collection of regulatory initiatives to oversee the processes by which platforms make rules and govern markets, rather than the services they host itself or the tools they use’³¹. It recognises that platforms are not themselves responsible for the content that users create or share, but that they do play a vital governance role by setting the policies which permit or disallow certain types of content and

³⁰ Victoria Nash and Mark Bunting, ‘A Policy Playbook for Platforms’ (2018) 46 *InterMEDIA* 30.

³¹ *ibid* 32.

behaviour, and further that platform architecture and code plays a role in shaping behavioural norms, affecting the visibility of content and monetising it through advertising. Procedural accountability then, holds platforms to account for these processes, policies and systems in line with principles of good governance³². On its own, procedural accountability does not imply a particular regulatory approach, and might be solely expressed through self-regulation (for example, voluntary publication of transparency reports, and moderation policies), or through co-regulatory or regulatory systems (for example the European draft Copyright Directive has procedural components, including transparency measures and standards for mechanisms of complaint and redress). A procedural accountability approach could go a long way to addressing many of the concerns set out in the OHWP, starting with the idea that in online activities, Internet users need to be empowered to protect themselves from the problematic behaviour of other users. A regulator might indeed be necessary to assess the adequacy of content moderation policies and processes in dealing with illegal content, for example; or an ombudsman could be asked to adjudicate whether platforms abide by their terms and conditions or moderation guidelines in user appeals. But vitally, the procedural accountability approach focuses on the messy detail of how platforms operate, rather than on whether a vague duty of care has been observed. The proposed 'duty of care' may in practice require procedural accountability on the part of platforms, but by choosing to target rhetorical simplicity over principles for good governance, it fails to provide a broad-ranging normative framework for platform governance that could truly set an international standard.

Revise and Resubmit? Priorities for the next iteration

Officials involved in the drafting of the OHWP have stressed that they are keen to learn from the consultation feedback, and that the next draft of the policy proposals may, as a result look rather different. With this in mind, it is worth summarising key areas for improvement:

First, any new regulatory proposals should not address both illegal and legal but harmful content. The former should be prioritised, with the expectation that any new interventions targeting technology companies would be grounded in an evidence base that identifies the biggest barriers to preventing, removing and investigating online illegal activity online. In an ideal world, this would lead not just to new measures aimed at improving co-operation and action by technology companies, but would also come with new money to ensure that law enforcement are fully resourced to tackle illegal behaviour online.

Beginning by tackling the most egregious content and harms would also enable a stepwise approach that would provide more time for an evidence-gathering exercise that could underpin future decisions as to whether harmful but legal online content needs more top-down regulation. Such a period of waiting would not come without responsibilities for technology companies. In the interim they would be expected to find ways of collaborating with academic researchers to enable the conduct of research investigating links between online content and experiences of harm.

Second, any country seeking international leadership in platform governance would be well advised to place human rights firmly at the centre of its proposals. It would signal, not just within the UK but

³² Mark Bunting, 'From Editorial Obligation to Procedural Accountability: Policy Approaches to Online Content in the Era of Information Intermediaries' (2018) 3 Journal of Cyber Policy 165.

to the wider world, that, internet regulation should not be misused to limit human rights and fundamental freedoms. In an era where our largest platforms now effectively function as public spaces, then at the very least the frameworks we are now establishing to regulate our online behaviours should not make it harder rather than easier for us to enjoy the full range of human rights³³. The development of such an approach by countries like the UK would help deter authoritarian regimes from misusing the models conceived and implemented in our liberal democracies.

Finally, as outlined above, a more coherent approach to developing policies that target online content and behaviour would embrace the complexities of platform governance, and focus on measures that hold platforms more accountable for their own procedures, processes and policies. An approach that embraced the principle of procedural accountability would rightly incentivise companies to ensure due process in all their dealings with users, which is vital when decisions are made that may ultimately limit freedom of expression, as well as other rights of participation, privacy and access to information. Vitally, the focus on holding companies accountable for their design choices as well as their explicit policies may also prove useful in tackling other platform problems such as algorithmic discrimination and data justice or concerns about lack of democratic accountability. If the UK could take a lead in establishing principles of good governance for the platform society, this would truly be a great gift to Internet users the world over.

³³ David Kaye, Report of the UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Human Rights Council 29th Session (United Nations 2018). <<https://freedex.org/a-human-rights-approach-to-platform-content-regulation/>> accessed 28 August 2019.