HOPE ~~HATE~~

A BETTER WEB

REGULATING TO REDUCE FAR-RIGHT HATE ONLINE

A BETTER WEB

REGULATING TO REDUCE FAR-RIGHT HATE ONLINE

HOPE NOT HATE'S RECOMMENDATIONS FOR REGULATING AGAINST ONLINE HARM

CONTENTS

INTRODUCTION

BE IT COVID-19 conspiracy theories shared in WhatsApp groups, campaigns of harassment by Twitter trolls, or the proliferation of far-right propaganda on YouTube, there is no doubt that harms perpetrated by extremists within the online world remain a pressing issue. HOPE not hate's research involves monitoring how extremists harm others online, and we are under no illusion as to the scale or breadth of the threat.

Today, major platforms like Facebook and Twitter are used by extremists for recruitment, propagandising at scale, disruption of mainstream debate, and the harassment of victims. Smaller platforms which have been co-opted, like Twitch and Discord, allow for further radicalisation and organising, as they are unable (or unwilling) to tackle extremists' abuse of their sites. Some platforms are even bespoke, structured to benefit those promoting extremism, as HOPE not hate's recent report into the video-sharing platform Bitchute highlighted.[1]

In response to this environment, civil society organisations are doing vital work pressuring tech companies to take greater action against these issues. However, it is increasingly clear that these dangers also require a deeper, regulatory solution.

The move to introduce internet regulation of this nature is a huge step. As Alex Krasodomski-Jones of Demos has rightly argued, "It is barely an oversimplification to characterise the current debate on internet regulation as a fight over the things people see, and the things they don't."[2]

To this end, the government's Digital Charter initiative, and the work that stems from it, is a welcome move. The Bill, if it resembles the white paper that preceded it, will aim to tackle harmful content and behaviours online in their entirety, from those that fall within HOPE not hate's remit, such as terrorist and hate content and activity, to issues as varied as child sexual exploitation and abuse, modern slavery, the sale of weapons and advocacy of self-harm. With such a breadth, there are concerns that the scope of the Bill will be too broad to be manageable, or that it will not be sufficiently detailed and could lead to hasty and flawed legislation. These are important considerations but they do not preclude working out how particular areas of online harms could be tackled. The nature of how online harms manifest in the digital realm through the far right's actions is one such complex area, and we have to ensure that the government's policy – in whatever form it takes – recognises and understands this.

Though aiming to remedy genuine harms, government regulation of our online lives also raises legitimate concerns over privacy and freedom of expression. We must address online harms whilst ensuring harms are not also inflicted through unfairly infringing on people's freedoms. HOPE not hate recognises the importance of this balancing act, and encourages a form of regulation of platforms that places democratic rights front-and-centre.

In a world increasingly infused with the web, the significance of this legislation cannot be overstated and it is undoubtedly the case that getting it right will take rigorous reflection. To that end, we encourage debate of the recommendations proposed here.

Finally, this is not just an opportunity to reduce the negative impacts of hostile and prejudiced online behaviour but also a chance to engage in a society-wide discussion about the sort of internet we do want. It is not enough to merely find ways to ban or supress negative behaviour, we have to find a way to encourage and support positive online cultures.

A NOTE ON THIS REPORT'S FOCUS

As an organisation that attempts to understand and respond to the extremist political landscape in the UK, we are well aware of the importance of online activity to extremists today. Though we campaign against all manner of extremisms, HOPE not hate's expertise and focus lies in tackling the organised far right.

As such, this report and our recommendations are particularly attuned to how legislation could undermine the online harms propagated by these actors.

At the same time, we recognise that far-right extremism does not exist in a vacuum and instead emerges from (and feeds back into) wider societal prejudices and inequalities. The activists and groups we campaign against target specific cohorts who are systemically on the receiving end of these prejudices and inequalities, especially women, members of ethnic minority groups, religious minority groups, and LGBTQ+ communities. One such manifestation of this is the continued, disproportionate abuse of members of these groups and others, such as people with disabilities, online.

To the extent that they can, these wider, systemic issues must be addressed in this legislation alongside efforts to curb extremism, and we encourage the government to listen to civil society recommendations on addressing these wider, societal factors online. In the UK, brilliant work is being done on this by groups including Glitch[3], the Antisemitism Policy Trust[4], Galop[5], The Fawcett Society[6], Tell MAMA[7], Leonard Cheshire[8] and many others.

2. WHAT ARE ONLINE HARMS?

THE CONCEPT of 'online harms' can be understood as referring to harms that occur on, or are facilitated by, the online world. An example of the former might be someone receiving threats via a direct message on social media. An example of the latter might be someone being radicalised by a video recommendation algorithm that suggests ever more extreme propaganda videos. Sometimes these harms can blur with our offline lives, for example when a terrorist group organises an attack using an encrypted private messenger, or if fake news about how to cure COVID-19 leads people to try dangerous home remedies.

The truth is that harms addressed by social media companies are defined by those companies. For some Holocaust denial is within scope, for others it is not. Whereas the UK authorities define harms outside the criminal law in other areas, for example in relation to the television we consume, harms are currently defined for us. That is not a tenable solution when dealing with the vast and ever-changing online world.

However, whilst many online harms are not clear cut, particularly those which can blur the line between legality and illegality such as trolling, it is unquestionable that some harms can and do occur on, or are facilitated by, the online world. At HOPE not hate, this has been clear to us from our research into how hate, division and fear are spread by extremists online.

The range of online harms the government wish to address extend far beyond HOPE not hate's remit – from Child Sexual Exploitation and Abuse (CSEA) to the sale of illegal goods – but a great many are central to what we campaign against. One of the government's priority concerns (alongside CSEA) is terrorist content and activity and, worryingly, the far right's use of the web to promote, plan and assist in terrorism is something HOPE not hate has increasingly witnessed in recent years.

The government acknowledged its initial list of harms is "neither exhaustive nor fixed", partly because it recognises that a "static list could prevent swift regulatory action to address new forms of online harm, new technologies, content and new online activities."[9] Some, such as those occurring on the dark web, are being addressed through separate government strategies. Of those it lists, the key areas for HOPE not hate are explained below:

TERRORIST CONTENT AND ACTIVITY

The government's white paper on online harms define this as terrorists' use of the internet "to spread propaganda designed to radicalise vulnerable people, and distribute material designed to aid and abet terrorist attacks. There are also examples of terrorists broadcasting attacks live on social media."[10]

EXTREMIST CONTENT OR ACTIVITY

The government's 2015 Counter-Extremism Strategy defines extremism, and by extension content or activity that can be considered extremist, as "the vocal or active opposition to our fundamental values, including democracy, the rule of law, individual liberty and the mutual respect and tolerance of different faiths and beliefs. We also regard calls for the death of members of our armed forces as extremist."[11]

HARASSMENT

The Equality Act 2010 defines harassment as unwanted conduct related to a protected characteristic which has the purpose or effect of violating the dignity of an individual, or creates an intimidating, hostile, degrading, humiliating or offensive environment for the individual.[12] Online harassment, however ,often extends beyond protected characteristics, for example when targeting MPs or journalists on the basis of their role (though amongst these, particular groups – e.g. female MPs or ethnic minority journalists – face a disproportionate amount of harassment, often on the explicit basis of their protected characteristics).

Online harassment can also blur into a number of other harms. Internet safety organisation

Glitch defines online abuse or harassment as "a catch-all term for various tactics and malicious behaviours online. This ranges from sharing embarrassing or cruel content about a person, impersonating, doxing and stalking, to the nonconsensual use of photography and violent threats. The purpose of harassment differs with every incidence, but usually includes wanting to embarrass, humiliate, scare, threaten, silence, extort or, in some instances, encourage mob attacks or malevolent engagements."[13]

HATE CRIME

The Crown Prosecution Service (CPS) defines hate crime, on the basis of a number of areas of UK legislation, as "Any criminal offence which is perceived by the victim or any other person, to be motivated by hostility or prejudice, based on a person's disability or perceived disability; race or perceived race; or religion or perceived religion; or sexual orientation or perceived sexual orientation or transgender identity or perceived transgender identity."[14]

INCITEMENT OF VIOLENCE

The CPS defines incitement – here referring to incitement of violence – as "incit[ing] another to do or cause to be done an act or acts which, if done, will involve the commission of [a violent] offence or offences by the other [person]" and the person inciting "intend[s] or believe[s]" that the other person will carry out this violent act or acts.[15]

TROLLING

HOPE not hate defines trolling as the act of being deliberately offensive or provocative online with the aim eliciting a hostile, negative, outraged reaction. Trolling is not covered explicitly in UK law but in practice often falls under laws that can be applied to online harassment and cyberbullying. For example, the Malicious Communications Act 1988 has been used to address cyberbullying. It states: "Any person who sends a letter, electronic communication or article of any description to a person that conveys a message that is indecent or highly offensive, a threat or false information. If the reason for that communication was to cause distress or anxiety to the recipient or to any other person, then the sender is guilty of an offence."[16]

INTIMIDATION

Intimidation, like trolling, often in practice falls under laws that address harassment. However, as the government's Committee on Standards in Public Life reported in 2017 in a review on the subject, it specifically concerns "words and/or behaviour intended or likely to block or deter participation, which could reasonably lead to an individual wanting to withdraw from public life."[17]

DISINFORMATION

The government's online harms white paper defines disinformation as "information which is created or disseminated with the deliberate intent to mislead; this could be to cause harm, or for personal, political or financial gain".[18]

VIOLENT CONTENT

The government's online harms white paper defines violent content as ranging from "content which directly depicts or incites acts of violence, through to content which is violent with additional contextual understanding or which is harmful to users through the glamorisation of weapons and gang life."[19]

3. WHAT IS THE ONLINE HARMS BILL?

IN APRIL 2019, the government published a white paper outlining their proposed policy for tackling online harms. In their most basic definition, they describe these as "behaviour[s] online which may hurt a person physically or emotionally. It could be harmful information that is posted online, or information sent to a person."

The policy aims to introduce a new statutory duty of care from companies in the scope of the regulation towards their users to protect them from online harms, compliance with which will be overseen by an independent regulator. The scope of companies in question is broad, addressing all which "allow users to share or discover user-generated content or interact with each other online", including "social media platforms, file hosting sites, public discussion forums, messaging services and search engines."[20]

The regulator – which the white paper suggests and the government has since re-iterated, could be Ofcom – will create codes of practice for companies to meet the duty of care and will be given enforcement powers in the case of failures to meet this duty. These may include the power to levy fines, make members of senior management liable, and possibly even require internet service providers to block platforms in the UK.

The regulator will not itself remove individual harmful pieces of content, instead its role will be to ensure companies are doing this adequately. The regulator will have the power to require annual transparency reports from companies on the prevalence of harms on their platforms and what measures are being taken to counteract these. These will be published online for the public, and the regulator would be able to demand further information, including the impact of algorithms on recommending content.

Companies' complaint services will have to abide by this duty of care and will be overseen by the regulator to ensure harmful content is responded to adequately, whilst an independent review mechanism will be created to ensure user concerns about removals from platforms, or lack thereof, can be addressed.

4. HOPE NOT HATE'S RECOMMENDATIONS

BETTER REGULATION of the web to undermine online harms is key, but, of course, is easier said than done. After the government ran a public consultation on the white paper between April-July 2019, they published the responses in February 2020, which drew attention to a range of blind spots and highlighted the competing interests at play in shaping this policy.

To help in the further shaping of this policy, below we outline a key approach to platform governance of online harms that the government should pursue, as well as some of the key ways far-right extremists relate to online harms.

1. TAKE AN INCLUSIVE APPROACH TO REGULATION

1.1 A regulator that includes the expertise of civil society

The online harms white paper, to some extent, pursues an independent approach to regulation through suggesting that the regulator could be Ofcom. However, distance should be ensured between government and the online harms regulator.

Moreover, the government should ensure that they adopt a path for platform governance that is truly democratic, inclusive and meaningfully involves the public and civil society, not just the private sector and the state.

- To this end, the government should explore the option of creating, in the first instance, a national SMC or other regulator either housed within Ofcom or set up as a separate, statutory corporation. This should be done with a view to joining this up with regional and international SMCs or other regulatory bodies in the future.
- The design of this regulator should be assisted through working with the leading researchers and experts on regulatory solutions to platform governance.
- The UK regulatory body should propose a duty of care from platforms which corresponds to UK law and international human rights law.

1.2 Inclusive creation of the codes of practice

We agree with the white paper consultation responses' call that members of affected groups "should be actively involved and consulted in designing safe products."[21] This speaks to the regulatory approach advocated throughout this report. Nonetheless, the white paper's brevity on this issue highlights a need for a more extensive understanding of this in the policy. For example, as Seyi Akiwowo, head of UK internet safety charity Glitch noted in July 2020, the policy still currently fails to acknowledge "those [with] multiple intersecting identities."[22] As such:

- The regulator should ensure that the codes of best practice for adhering to the proposed statutory duty of care are drawn up in an inclusive manner, including not only the state and companies in the scope of the

legislation, but also civil society and members of groups known to be marginalised online, as well as those targeted by extremists.

- The regulator should ensure that codes of practice take into account the increased likelihood of individuals experiencing online harms when they have multiple intersecting identities that are targeted by extremists. For example, if someone is both black, female and a figure in the public eye.
- The codes of practice should be regularly reviewed through regulatory meetings with representatives from the public, civil society, the state and the companies in the scope of the policy.

2. ENSURE FREEDOM OF SPEECH NOT FREEDOM TO HARM

The companies in the scope of the online harms legislation occupy central roles in the public sphere today, providing key forums through which public debate occurs. However, it is vital that they ensure that the health of discussions is not undermined by those who spread hate and division.

- At present, online speech which causes division and harm is often defended on the basis that to remove it would undermine free speech. However, in reality, allowing such speech to be disseminated only erodes the quality of public debate, and causes harms to the groups such speech targets. This defence, in theory and in practice, minimises free speech overall. This regulation instead should aim to *maximise* freedom of speech online for more people, including those from minority backgrounds whose speech is consistently marginalised online and elsewhere. This principle should be front-and-center of the government's public information campaign surrounding this bill, as it otherwise stands to be misconstrued as an infringement upon free speech. For this reason, any such campaign also ought to be clear about what the regulator will and will *not* be able to do, so that it cannot be misrepresented.
- The regulator should have powers to look at specific cases of disputed online harms which are particularly high-profile or serious, so-called 'super-complaints'. This would allow wider debates over the (de)platforming of high profile extremists, such as Stephen Yaxley-Lennon (AKA Tommy Robinson), to be carried out more carefully. Conversely, complaints raised against deplatforming could be brought by more fundamentalist free speech organisations or activists. This would be particularly beneficial for setting a precedent to platforms when it comes to their moderation of novel or more complex harms.
- There ought to be a measure of who is *not* on a platform to highlight and understand marginalisation through harmful and divisive speech. Tracking engagement with users permission, and following up with those who have deleted their accounts or become inactive to ask why, will give a clearer picture of who is being driven away from a site by virtue of other users behaviour.

3. ENSURE PRIVACY WHILST ENSURING PROTECTION FROM ITS ABUSE

Privacy is a vital freedom but it can be abused by those who orchestrate harms covertly. Addressing the issue of online harms found in or originating from private communications, one recommendation has been that the regulator require companies to investigate groups online which surpass a maximum member threshold, beyond which a conversation is no longer considered private.

However, this raises concerns over arbitrariness and, more importantly, doesn't address the issue squarely on the basis of the harm being caused itself but rather on a proxy to this, namely membership of a group. It would also mean small but extreme private groups would be overlooked, and also does not clearly delineate on the nature of the harms that would be investigated.

- An alternative recommendation for investigating private communications used by those propagating online harms is that those communications investigated are those where there is a "strong evidence basis" that online harms are resulting from them, as the Bonavero Institute of Human Rights at Oxford University recommend.[23]
- Evidence should be brought to the regulator by the public and by civil society groups who monitor extremist activity.

- Where there is a strong evidence basis that the harms being propagated are illegal, investigations should be carried out by the police and civil society groups and the public should bring this evidence to them in the first instance.
- Where the legality of the harms being propagated is unclear, the responsibility should rest with the platform to investigate once a complaint is made.

Some have argued that the presence of a blocking feature on a private service, enabling users to block content from other individuals, is sufficient for tackling harm. This fails to address the harm of radicalisation on an individual. Leaving the responsibility to judge whether terrorist and other extremist content is harmful in this way to the person potentially being radicalised is clearly ineffective, and would not prevent many vulnerable people – including children, the priority cohort for the government's policy – from falling prey to propaganda.

4. BE CONSISTENT IN TACKLING KNOWN EXTREMISM

It is important that companies tackling online harms view individuals, groups and organisations known to promote extremism in a manner that goes beyond a single event and beyond their platform. To this end:

- As part of their Duty of Care, the platforms should be required to remove or give warnings to accounts on the basis of their content that is reported as violating their terms of service or which breaks the law, rather than just removing the content alone but allowing the account to stay up.
- The regulator should require companies to take into account the actions and behavior of individuals outside of their platform when deciding whether to remove or give warnings to them. At present, some people engage in hate speech and/or violence in the real world but moderate their tone on social media platforms to avoid moderation and deplatforming. The result is that major platforms are used to organise events and movements by individuals who engage in extreme behaviour elsewhere.
- The regulator should investigate coordinated mass activity flagged as causing illegal harms, or with the potential to cause these. Evidence for this should come from the public and civil society organisations, and particular attention should be given to evidence that it is indeed coordinated, i.e. that it originates from a certain individual or a group wishing to cause harm. If this is established, the regulator should have power to require platforms to disrupt the activity.

5. DESIGN AGAINST HATE

Extremists' abilities to perpetrate online harms are often exacerbated by the design of sites by companies proposed to be in the scope of this regulation, and different kinds of platforms present different opportunities for extremists. Major platforms like Facebook and Twitter allow recruitment, propagandising at scale, disruption of mainstream debate, and the harassment of victims. Co-opted platforms, like Twitch and Discord, allow for further radicalisation and organising, as they are smaller and unable to tackle extremists' abuse of their sites. Some platforms are bespoke, structured to benefit those promoting extremism, as HOPE not hate's recent report into the video-sharing platform Bitchute highlighted.[24]

Attention to these differences is necessary for the online harms policy to effectively tackle online extremism. Limiting the focus to just sites set up for unlawful purposes misunderstands the nature of online harms such as terror and hate speech content. A great deal of such content appears regularly on websites not set up for unlawful purposes, such as the notorious image board 4chan. With these considerations in mind, the regulator ought to:

- Build into the duty of care prohibitions and recommendations on platform technology design. Prohibitions could be against designs known to cause harm, for example, particular recommendation algorithms known to lead to ever more extreme content. Recommendations could include best practice on platform technology design, and this should be open to revision given further research.
- Where an investigation into design does not clearly fall under the remit of the police – i.e. it lies within the grey area of il/legal harms – the regulator should examine evidence from the public and civil society groups that a platform is encouraging harms through its design.

- If it is shown to be the case that a platform is knowingly encouraging illegal harms, the police should require that its IP address be banned, require the ban of cross-platform sharing of links to the platform in question, and ensure that relevant criminal proceedings are brought against the platform, including any senior management liability.
- If it is shown to be the case that a platform is not encouraging but is nonetheless inadequately handling the perpetration of online harms on their site, they should be flagged by the regulator as a risk to vulnerable users. Relevant authorities and NGOs that work with vulnerable people should be notified of this and the platform should in the first instance work with the regulator to ensure recipients of online harms on their platform are shielded from this. If such a platform fails to address these harms they should be fined, and should be given escalating penalties for continued failure, potentially resulting in criminal proceedings.
- Require that platforms ensure that users vulnerable to radicalisation are less likely to be led towards extremists online. The duty of care established by the regulator should ensure that companies “minimise risk by design and default”, as children’s online safety charity 5Rights highlight. This would mean holding companies to account for features such as “their recommendation algorithms, user journeys, age-assurance mechanisms, and default settings”.[25]
- Require companies to prioritise counteracting what extremists primarily use their platforms for, e.g. Twitter should prioritise counteracting the spreading of disinformation, whereas Discord should prioritise the use of small, private groups for radicalisation.
- Require platforms which attract larger numbers of visitors to match this with a greater number of moderators, and ensure that a proportion of these engage in internal searching for illegal harms on the platform, rather than waiting just for external reporting of these by users. The proportion of such moderators should be determined on a tiered basis across platforms relative to the number of visitors.
- Require platforms to ensure support for their moderation teams that spend large amounts of time engaging with extreme content.
- Require companies to invest in technologies which can be used for conflict resolution on their platforms between users.
- Require companies that fall under the scope of the legislation but which are based outside the UK to appoint a nominated representative for the UK.

6. NO EXCUSES ON ONLINE HARM GREY AREAS

Understanding the changing nuances of hate spread online is essential for tackling it. The fact that some online harms are complex (such as disinformation campaigns), novel to those not on the receiving end (such as the experiences of hate directed at marginalised groups), or lie at the border of il/legality (such as when trolling veers into harassment), should not be used as a defence by companies for not adequately ensuring users are not exposed to them. To this end:

- The regulator should ensure that the duty of care requires companies to stay up to date on the changing nature of online harms, and continue to reflect this in their platform design and policies. The regulator can lead in developing best practice on this by convening civil society, the public, the state and platforms to share knowledge.

7. PROVIDE WIDER EDUCATION ON ONLINE HARMS

Whilst this legislation will be a landmark change in the tackling of online harms in the UK, it cannot change online behaviour and culture by itself. Moreover, by their nature online harms are complex and open to change. Given these factors, continued education is key. To this end:

- The regulatory body should engage in recurring public information campaigns to reflect the changing nature of online harms, or emergence of new harms. This should be not just for the public at large, but also for arms of the state including the police, judiciary and other essential services for the body democratic.
- As part of the duty of care, and to complement the regulators efforts,

companies in the scope of the policy ought to ensure their users are informed about online harms.

- Civil society initiatives aiming to educate the public on the nature of online harms should be supported, through working with the regulator as well as bidding for government grants available for such work and by cooperation directly with companies in the scope of the policy.
- All initiatives aimed at educating around online harms – be they from the regulator, the state, platforms, or civil society – ought to encourage a proactive approach. This reflects a point raised in the white paper consultation by organisations working against violence against women and girls. As they highlight: "education for online safety should focus not only on behaviours to adopt, but also on discouraging adoption of negative behaviours."[26] In this way it would help "prevent the content from coming online in the first place".[27]
- All initiatives aimed at educating around online harms – be they from the regulator, the state, platforms, or civil society – should also promote awareness about how to engage in discussions online without marginalising or silencing others and in so doing, undermining their freedom of expression also.

5. CASE STUDY: LEARNING FROM INTERNET REGULATION IN GERMANY

ON 1 JANUARY 2018, the Network Enforcement Act (NetzDG) came into effect in Germany. The law was brought in after recognition that social media companies had inadequately dealt with far-right anti-refugee sentiment online following Germany's acceptance of one million refugees in 2015. Amongst other things, the law requires social networks with two million or more users to take down or block reported criminal content within 24 hours of receiving a report (when its legal status is uncertain, they have seven days). Companies can also receive up to €50 million in fines if there are systematic infringements of the requirements, and they must also publish transparency reports regarding compliance with NetzDG. Considering this law is instructive for the wider question in the background of this report, of how we ought to govern social media and other internet platforms.

A key concern about the law was that it would encourage people not to express opinions for fear of removals (known as the 'chilling' of freedom of speech), particularly if platforms overreacted to the law (known as 'overblocking') by taking down content falsely reported as illegal, something which could be abused by malicious actors. However, as Professor Wischmeyer at the University of Bielefeld has highlighted, issues with the lack of detail and ability to compare the transparency reports means they "can neither confirm nor refute the 'censorship' or the 'over-blocking' claim", but they do "demonstrate that some of the fears associated with the law have been clearly exaggerated", since "the numbers for [NetzDG] blockings are very low".[28]

The deeper underlying issue researchers on NetzDG have raised following the reports is that it has delegated power to platforms, by virtue of leaving the decision of content's lawfulness to the discretion of moderators; arguably, effectively "privatising" the judiciary in this context.[29] Moderators already have stressful work conditions, so if faced with the task of assessing illegality in a short time span or risk their employer a vast fine, it is unsurprising that the transparency reports indicate that "the community standards" not national law "remain the principal denominator in assessing the legality of content."[30] It is also unsurprising, therefore, that some suspect platforms have in some cases deliberately made NetzDG violation reporting mechanisms not user-friendly.[31]

Pending amendments to the law aim to address this and other key issues raised with NetzDG, such as disclosure of content removed for court uses, strengthening user rights through appeal procedures, and making the transparency reports more detailed. There is no addressing, however, of the underlying issue of the balance of power between the state (and users) and platforms, though there has been a call by German anti-extremism NGO Amadeu Antonio Stiftung (AAS) for a "German internet forum with equal involvement of internet companies, government and parliamentary representatives, as well as representatives of civil society" to "encourage closer cooperation".[32] This is meant to encourage better negotiation and so in practice could mean encouraging better self-regulation by platforms, rather than necessarily calling for increased regulation by the state.

In terms of achieving a safer web, the balancing of power here – between users, the state and companies - is far from straightforward, but lies at the heart of what must be reckoned with in all online harms legislation. As Heidi Tworek wrote for The Centre for International Governance Innovation in 2019, platform governance can involve degrees of government intervention ranging from hands-off, business self-regulation to outright statutory regulation.[33]

Between these are various options for 'co-regulation' between the state, the private sector and the public.

Until now, self-regulation has been the *de facto* stance around the world, but clearly relying on companies to do better, however much they are cajoled, is not sufficient. Equally, strict, penalising statutory regulation like NetzDG can backfire in ways and can lead to the fear, and the reality, of freedoms being infringed. Co-regulation attempts to tread a middle path, allowing platforms some freedom to tackle online harms as they see fit as long as these efforts meet certain standards, and ensuring these are met by making both the regulator and the regulated accountable to the public.

One increasingly popular proposal in this realm is that of 'social media councils' (SMC). SMC is defined by the freedom of speech and freedom of information charity Article 19 as "a multi-stakeholder accountability mechanism for content moderation on social media", which would aim to provide an "open, transparent, accountable and participatory forum to address content moderation issues on social media platforms on the basis of international standards on human rights."[34]

In essence, SMCs are independent, transparent bodies which could adjudicate on and/or advise/mediate between the private sector and government on online harms. They could operate on a regional or national level and potentially work on an international scale too in partnership with SMCs elsewhere. SMCs could offer a democratic solution to moderation, drawing in representatives from across the board as AAS called for, whilst perhaps going beyond mere encouragement to having some power to enforce decisions made on social media regulation.

However, there is a long way to go for consideration of SMCs and similar proposals, as there are "a wide range of organizational structures and precedents to consider, with the format, jurisdiction, makeup, member selection, standards, and scope of work subject to debate".[35] It is beyond the scope of this report and HOPE not hate's remit to suggest the best solution of this kind, but as an intermediary between the NetzDG model of strict statutory regulation and currently inadequate platform self-regulation, a co-regulatory option of this sort, points to a promising path for platform governance and should be the approach taken by the UK government's online harms policy.

ENDNOTES

[1] Davis, G. July 2020. Bitchute: Platforming Hate and Terror in the UK. UK. HOPE not hate Charitable Trust. https://www.hopenothate.org.uk/wp-content/uploads/2020/07/BitChute-Report_2020-07-v2.pdf

[2] Krasodomski-Jones, A. October 2020. Everything in Moderation: Platforms, Communities and users in a healthy online environment. https://demos.co.uk/project/everything-in-moderation-platforms-communities-and-users-in-a-healthy-online-environment/

[3] Glitch. 2019. ‘Response to Online Harms White Paper’. UK. Available at: https://fixtheglitch.org/2019/04/08/response-to-online-harms-white-paper/

[4] Antisemitism Policy Trust. ‘Policy Briefings & Reports’. UK. Available at: https://antisemitism.org.uk/research-reports/

[5] Hubbard, L. 2020. Online Hate Crime Report 2020: Challenging online homophobia, biphobia and transphobia. UK. Available at: http://www.galop.org.uk/online-hate-crime-report-2020/

[6] The Fawcett Society. 2019. ‘Online Harms White Paper: Consultation (Fawcett Society Submission)’. UK. Available at: https://www.fawcettsociety.org.uk/online-harms-white-paper-consultation-fawcett-society-submission

[7] Tell MAMA. ‘Reports’. UK. Available at: https://tellmamauk.org/category/reports/

[8] Leonard Cheshire. 11 May 2019. ‘Online disability hate crimes soar 33%’. UK. Available at: https://www.leonardcheshire.org/about-us/press-and-media/press-releases/online-disability-hate-crimes-soar-33

[9] HM Government. April 2019. Online Harms White Paper. UK. 30. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf

[10] Ibid. 5.

[11] HM Government. June 2011. Prevent Strategy. UK. 107. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/97976/prevent-strategy-review.pdf

[12] HM Government. October 2010. Equality Act 2010: Chapter 15. UK. 26-27. https://www.legislation.gov.uk/ukpga/2010/15/pdfs/ukpga_20100015_en.pdf

[13] Glitch. ‘Online abuse explained’. UK. Available at: https://fixtheglitch.org/online-abuse/#:~:text=Online%20Abuse%20is%20a%20catch,of%20photography%20and%20violent%20threats.

[14] HM Government. October 2016. Hate Crime: What it is and how to support victims and witnesses. UK. 2. https://www.cps.gov.uk/sites/default/files/documents/publications/Hate-Crime-what-it-is-and-how-to-support-victims-and-witnesses.pdf

[15] Crown Prosecution Service. 21 December 2018. ‘Inchoate offences’. UK. Available at: https://www.cps.gov.uk/legal-guidance/inchoate-offences [Accessed 18.9.20]

[16] HM Government. July 1988. Malicious Communications Act 1988. UK. Available at: https://www.legislation.gov.uk/ukpga/1988/27 [Accessed 18.9.20]

[17] HM Government. December 2017. Intimidation in Public Life: A Review by the Committee on Standards in Public Life. UK. 26.

[18] HM Government. April 2019. Online Harms White Paper. UK. 22. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf

[19] Ibid. 67.

[20] HM Government. April 2019. Online Harms White Paper. UK. 8. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf

[21] HM Government. February 2020. Online Harms White Paper – Initial consultation response. UK. 51. http://data.parliament.uk/DepositedPapers/Files/DEP2020-0111/Online_Harms_White_Paper-Initial_consultation_response.pdf

[22] Akiwowo, Seyi (@seyiakiwowo). 12 July 2020. 12:08pm. ‘4. All political parties should have a policy on the Online Harms Bill that is intersectional. The bill will be the first piece of social media regulation and currently fails to acknowledge women and those multiple intersecting identities.’ https://twitter.com/seyiakiwowo/status/1282270703585296385?s=20

[23] Theil, S., Butler, O., Jones, K, Moynihan, H., O’Regan, C., Rowbottom, J. UK. 1 July 2019. ‘Response to the public consultation on the Online Harms White Paper’. Bonavero Report No.3/2019. 8. https://www.law.ox.ac.uk/sites/files/oxlaw/bonavero_response_online_harms_white_paper_-_3-2019_0.pdf

[24] Davis, G. July 2020. Bitchute: Platforming Hate and Terror in the UK. UK. HOPE not hate Charitable Trust. https://www.hopenothate.org.uk/wp-content/uploads/2020/07/BitChute-Report_2020-07-v2.pdf

[25] 5Rights Foundation. May 2020. ‘Home Office preparedness for Covid-19 (online harms). UK. 5. https://5rightsfoundation.com/uploads/final-5r-response-to-hasc-consultation-on-covid-19.pdf

[26] HM Government. February 2020. Online Harms White Paper – Initial consultation response. UK. 53. http://data.parliament.uk/DepositedPapers/Files/DEP2020-0111/Online_Harms_White_Paper-Initial_consultation_response.pdf

[27] Ibid. 54.

[28] Wischmeyer, T. 2019. ‘”What is Illegal Offline is Also Illegal Online” – The German Network Enforcement Act 2017’. In Fundamental Rights Protection Online: The Future Regulation of Intermediaries (Eds. Petkova, B., Ojanen, T.). Edward Elgar Publishing. UK. 20.

[29] Heldt, A. 2019. ‘Let’s Meet Halfway: Sharing New Responsibilities in a Digital Age’. Journal of Information Policy (9). Penn State University Press. USA. 342.

[30] Schmitz, S., Berndt, CM. 2018. ‘The German Act on Improving Law Enforcement on Social Networks (NetzDG): A Blunt Sword?’. Available at SSRN: https://ssrn.com/abstract=3306964.

[31] Tworek, H., Leerssen, P. 27 February – 3 March 2019. ‘An Analysis of Germany’s NetzDG Law’. First session of the Transatlantic High Level Working Group on Content

Moderation Online and Freedom of Expression. UK. 5. https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf

32 Rafael, S., Ritzmann, A. 2019. 'Background: The ABC of hate speech, extremism and the NetzDG'. In Hate Speech and Radicalisation Online: The OCCI Research Report. Baldauf, J., Ebner, J., Guhl, J. (Eds.). Institute of Strategic Dialogue. UK. 16.

33 Tworek, H. 'Social Media Councils'. 28 October 2019. The Centre for International Governance Innovation. Canada. https://www.cigionline.org/articles/social-media-councils

34 Article 19. 11 June 2019. 'Social Media Councils: A Consultation'. Article 19. www.article19.org/resources/social-media-councils-consultation/

35 Ness, S., Schaake, M. 13 February 2020. 'Co-Chairs Report No.3: The Bellagio Session'. Third session of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression. UK. 8.

HOPE not hate
PO Box 61382
London N19 9EQ
United Kingdom
t: +44 (0)20 7952 1181
e: office@hopenothate.org.uk
w: hopenothate.org.uk