OPEN RIGHTS GROUP

Response to Consultation on the Online Harms White Paper July 2019

Contents

0. Summary of principles 1

1. Introduction 1

2. Framing the regulatory narrative 2

3. Taking a rights-based approach 5

4. Issues with risk and harms models and the proposed duty of care 6

5. A call for co-regulation 8

6. Regulation should be evidence-based and process-driven 10

7. Need for counter-notice procedure in cases of alleged illegality 10

8. Supporting democracy and criminal justice 13

9. Lawful content is lawful content 14

10. Consider the free expression impact of liability and enforcement 15

11. Transparency and accountability 17

12. Accountability for State actors 20

13. Use regulation to build public trust 22

0. SUMMARY OF PRINCIPLES

- Rights-based regulatory framing
- Right to publish legal content
- Accuracy as well as speed
- Access to effective redress
- Equality of arms
- Avoid measures that compel takedown of legal content

1. INTRODUCTION

1.1 Open Rights Group (ORG) is a UK-based digital campaigning organisation working to protect fundamental rights to privacy and free speech online. With over 3,000 active supporters, we are a grassroots organisation with local groups across the UK.

1.2 ORG has actively engaged with the government's proposals for online regulation since the Internet Safety Strategy in 2017. The following comments have been developed through a long period of reflection, report writing and engagement with different stakeholder groups. Our main reports underpinning this response are:

a. Internet Regulation, Parts I[1] and II[2]
b. Blocked: Collateral Damage in the War against Online Harms[3]
c. DNS Security: Getting It Right[4]

1.3 In recent months, we have met with tech industry representatives, children's rights and safety groups, academics, lawyers and think-tanks, government departments and regulatory bodies to discuss the issues raised in the White Paper. We also co-hosted an interactive multi-stakeholder workshop bringing together all these different perspectives in a discussion setting to explore difference and find consensus.[5] Our thinking in respect of the White Paper

[1] Open Rights Group, *UK Internet Regulation Part 1*, December 2018 <https://www.openrightsgroup.org/assets/files/pdfs/reports/Internet_Regulation_Part_I_Internet_Censorship_in_the_UK_today-web.pdf>

[2] Open Rights Group, *UK Internet Regulation Part 2*, June 2019 <https://www.openrightsgroup.org/assets/files/pdfs/reports/ORG_Regulation_Report_II.pdf>

[3] Open Rights Group, *Collateral Damage in the War against Online Harms*, April 2019 <https://www.openrightsgroup.org/assets/files/reports/report_pdfs/top10vpn-and-org-report-collateral-damage-in-the-war-against-online-harms.pdf>

[4] Open Rights Group, *DNS Security: Getting It Right*, June 2019 <https://www.openrightsgroup.org/assets/files/reports/report_pdfs/ORG_DNS_Security_Report_.pdf>

[5] Victoria Nash, Oxford Internet Institute, *Internet Regulation and the Online Harms White Paper: Stakeholder Workshop Summary*, 1 July 2019 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3412790>

proposals has been refined, enriched and sharpened as a result of having these diverse conversations.

1.4 We welcome this opportunity to respond to the White Paper. Our narrative response opts not to address specific questions but rather deals with issues and concerns more holistically, and comments supplement verbal contributions we have made at the topic-focused roundtable meetings hosted by DCMS. References to parts of the White Paper are included in square brackets [x].

1.5 We note that as the government continues to develop this regulatory proposal, all relevant stakeholders, including civil society and smaller/niche platforms, should be fully engaged and able to participate in the design and implementation of any measures which are finally adopted. If the government wishes to develop a truly effective system of regulation, it needs to listen to expert voices, engage with practical tech and legal realities, and be prepared to compromise on its own ambitions and desires.

2. FRAMING THE REGULATORY NARRATIVE

2.1 It is critical to emphasise at the outset that any model of online content regulation will always ultimately bite on the end user. It is individual citizen speech that will be curtailed - or potentially empowered - by such regulation, and the government must keep this impact on people firmly in mind as it develops its policy proposals. Make no mistake, the model of regulation the White Paper envisages is a censorship regime; the government should be treading much more carefully to avoid setting up a system that results in widespread suppression of lawful online content.

2.2 We recognise and acknowledge that for certain categories of individual, the Internet can be an unwelcoming and often hostile place. Nonetheless, in our view, the potential negative impacts that user-generated content can have on individuals, groups or even society is not the right starting point for regulation; rather, we should look to what companies are doing around that content, and in particular, how they are distributing and monetising it. Social media companies are private entities operating for commercial profit which ultimately make decisions based not on societal good but on their own financial interests. It is their data-driven business model, powerful control over citizen speech and operation within an online environment whose unique characteristics affect and influence the reach and impact of content, activity or behaviour, that together justify policy intervention. This is a different starting point to the White Paper, and produces different potential interventions.

2.3 The White Paper's proposed regulatory scope [4.1-4.3] is unrealistically vast. Search engines, cloud service providers and independent product review sites - to take but a small chunk of the array of services in scope - have very different operational functions and regulatory needs. In order to be successful, this scheme needs to be narrower and more focused. We urge the government to return to a more realistic and pragmatic vision of regulating social media platforms that handle the publication of very large volumes of user-generated material. This would also have the advantage of being a more natural fit to address the government's concerns driving online content regulation.

2.4 The government is right to rule out attempting to regulate private communications [4.7]. We would welcome greater clarity on how the lines between public and private are proposed to be drawn, particularly in view of international human rights law considerations and restrictions. We note that numerical indicators alone are unlikely to suffice.

2.5 In looking to regulate social media platforms, it must be acknowledged that platforms actively court user attention. The longer users spend on platforms, the more data about them is accrued and the more opportunities for monetisation magnify. To maximise user engagement, platforms seek to ensure that "interesting" content reaches people more quickly, which sets up the possibility that untruthful, exaggerated or overly-emotive content may be more likely to "succeed" in online spaces. As long as this attention-and-data business model thrives, interventions aimed at regulating particularly extreme content are likely to be relatively limited in their impact. More holistic regulation is therefore needed to address how online companies collect and process user data. If data protection law is enforced effectively, that could have a positive knock-on effect on online content distribution and moderation processes, as companies generally begin to behave in a more responsible way towards their users.

2.6 The ultimate aim of Internet regulation should be to ensure and support a digital environment that protects and respects human rights. We fully endorse the call of the UN Special Rapporteur on the right to freedom of opinion and expression, Professor David Kaye, for "smart regulation"[6] primarily focused on increasing and improving online companies' transparency and accountability. In our view, a systemic process focus would be the right approach, including audit functions to assess overall company performance around content distribution and moderation.

[6] Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, *Report: A Human Rights Approach to Platform Content Regulation*, 6 April 2018, A/HRC/38/35 <https://documents-dds-ny.un.org/doc/UNDOC/GEN/G18/096/72/PDF/G1809672.pdf?OpenElement >

2.7 To the extent that regulation does address individual pieces of content, it should focus on ensuring that companies (a) expeditiously detect and remove unambiguously illegal or unlawful content (with the caveat that groundless monitoring of content is not acceptable), and (b) set human-rights-compliant terms & conditions around lawful content - and uphold these consistently and transparently, including providing equal access to user redress mechanisms.

2.8 We note the government's ambition of creating a world-leading Internet regulation framework but urge caution. Social media platform experience is politically and culturally context-sensitive. The UK has strong media plurality, generally effective justice processes and well-established democratic institutions; circumstances which are not universally guaranteed. We strongly recommend that the government focus on developing a precise and effective regulatory scheme within the UK: if this is successful it will organically be replicated overseas, and the international duplication which is sure to occur will be the stronger. The government must be cognisant that instituting a broad regulatory scheme at home will increase risks of citizen oppression abroad.

2.9 The government also states that it wants to ensure the same rules online as offline [2.7]. Whilst this is a positive policy goal, it is not achieved by the current proposals. These imply a differential privatisation of justice online, with the assumption that corporate policing must replace public justice for reasons of convenience. De facto, civil and criminal legal standards continue to define the limits of speech offline but corporate terms and conditions (that already exclude some legal material) become the arbiter of what is acceptable speech online, reinforced by a regulatory requirement to exclude material that, offline, would be legal. Although companies are not being asked to define what is legal or illegal, regulation pressures companies to rigorously enforce terms and conditions where material may be unlawful - which pushes them to increase removal of legal material. This dangerously blurs the line between whether the policy objective is to remove unlawful material, or whether it is to remove more unwanted material - even where this is legal.

2.10 The White Paper's harms model also tends towards removing material that is morally objectionable, in violation of universal human rights standards. It is easy to point at negative user conduct on platforms and conclude that because this takes place within a defined space it is the responsibility of the platform to stop and even prevent, but this argument conveniently sidelines the principle of online-offline equivalence. Legitimate free expression is protected everywhere it occurs, and what is legal offline must remain legal online.

3. TAKING A RIGHTS-BASED APPROACH

3.1 To be lawful, precise and effective, any regulatory scheme must be explicitly rooted in the international human rights framework. This provides an objective, well-established standard capable of holding both corporate entities and States to account. We urge the government to redevelop its policy proposals so that the regulation has an explicit rights framework.

3.2 Whilst all rights are interconnected and interdependent, the right to free speech is evidently of critical relevance here. Laws protecting human rights apply equally online as offline; consequently, regulation must comply with the legality, legitimacy and necessity provisions established in Article 19 ICCPR and other international laws and treaties.

3.3 The White Paper pays little more than lip service to free speech; we urge the government to focus more practically on how this fundamental right will actually be protected in its proposals. Protecting freedom of expression - which includes the right to access information - must infuse how the government develops, implements and enforces its legislative and regulatory scheme. It is trite law that the right to freedom of expression protects speech that is offensive, disturbing and shocking, and that restrictions must be defined by law, serve a legitimate purpose and be the least restrictive means necessary to achieve that purpose. Any policy intervention that curtails speech must be defined and limited by precise terminology: imprecise language risks dangerous overreach. These legal limits must more explicitly guide the government in formulating its regulatory policy.

3.4 Regulation should further encourage companies to adopt and implement the UN Guiding Principles on Business and Human Rights.[7] These set out principles of due diligence, transparency, accountability and remediation, and would commit companies to implementing human rights standards throughout their product and policy operations. We would welcome incorporation of these principles into any regulatory framework so that they become directly enforceable. This would also fulfil the aim of targeting regulatory responsibilities at companies, rather than individual users.

3.5 The right to privacy additionally has relevance in this regulatory process and currently too little attention is given to this. Privacy particularly comes into play in relation to content monitoring requirements. The E-Commerce Directive 2000 prohibits general monitoring by platforms of user-generated content.[8] This prohibition is a sensible precaution against the disproportionate use of

[7] United Nations, *Guiding Principles on Business and Human Rights*, HR/PUB/11/04 <https://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf>

[8] Directive 2000/31/EC on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (Directive on electronic commerce), 8 June 2000, at [15] <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32000L0031&from=EN>

technology to establish suspicionless surveillance for a variety of possible behaviours; a tempting but unwise policy direction. However, general monitoring is the obligation presently created by the White Paper: this puts in place proactive duties on platforms to monitor for certain types of content, and the idea that “specific monitoring" [3.12] can be undertaken without general monitoring is a fallacy. The duty of care and codes of practice model of regulation in the White Paper pushes online platforms towards increased monitoring of user content in order to identify problematic material and ensure it is not seen by those to whom it could be harmful (however this is defined) - including through upload filters and blocks. Any pre-emptive monitoring is inconsistent with the right to privacy, and will lead to increased censorship, including by having a chilling effect on users’ own willingness to generate content. We recommend that, rather than leaning into monitoring, legislation in this process or separately should make the general monitoring prohibition explicit.

3.6 Regulation should further promote non-discrimination in decision-making, both human and algorithmic.

4. ISSUES WITH RISK AND HARMS MODELS AND THE PROPOSED DUTY OF CARE

4.1 Although a duty of care [3.1] is at the heart of the White Paper, this is a poor conceptual fit for addressing the societal challenges of social media platforms and should <u>not</u> form the basis of regulation. We understand why the duty of care idea is attractive but we have severe concerns with the application of the model to online speech and urge the government to seriously reconsider using it as the basis of regulation. We note that our concerns are echoed by many other organisations.

4.2 Duties of care are based on the notion of risk management. They can operate well in situations where an owner of a physical space or the provider of a tangible service might directly create risks for individuals interacting with that space/service if they are not sufficiently careful. However, these are normally risks which the owner/provider can directly control. A club or bar might owe a duty of care to their customers not to leave hazardous items strewn about where patrons might trip and injure themselves, but they would not normally have a direct responsibility for any speech or act done by one of their customers to another - except to the extent that e.g. a bar brawl reaches a criminal threshold, in which instance they might be obligated to call the police. Similarly, in the online space, it would be a strange outcome for an Internet platform to be placed under a duty to ensure that one user does not by their speech cause harm to another when that platform has no direct control over either individuals' speech - or even over their choice to interact with one another.

4.3 Focusing on risk and harm naturally produces models of content removal, rather than promoting fair, necessary and accurate actions and decisions. This makes it inappropriate for most kinds of speech-based situations. A risk-based approach could be appropriate in the case of clearly criminal content and activity. However, where risks are harder to discern, only apply to certain individuals or groups, or are wider societal risks rather than direct personal risks, the case for intervention becomes harder to make and the potential for overreach becomes greater. The White Paper fails to explicate any detail on how the duty of care is to operate, but if it is drawn broadly, including by extending the applicable definition of harm to include harm to individuals, vulnerable groups and society at large, the risks to free expression are particularly acute.

4.4 Additionally, in situations of speech, many possible issues of risk, such as harassment or bullying, may involve multiple parties with potentially different views of their behaviour who would each be owed a duty of care by the platform provider. Online behaviour may also be tangential to some offline behaviour where the real risks play out. A duty of care approach may find it very hard to address this, as it may be unreasonable to expect a platform to owe a duty of care relating to activity that takes place beyond its confines.

4.5 In any risk-based approach, the key question is the manner in which risk is established, the kinds of risk addressed, to whom the risks apply, the potential mitigations identified and the proportionality and speech impacts evaluated. There is no detailed discussion of these aspects in the White Paper, beyond asserting that they will need to be dealt with by the regulator. It is notable that in various meetings we have attended around the White Paper, the government so far has only said what this duty of care is *not*, rather than what it *is*. This indicates a lack of conceptual clarity, which will frustrate any reliance on it as the model.

4.6 Even if a risk model could be said to work for speech content, delineating specific harms to be addressed is an inherently problematic enterprise. The "harms" listed in the White Paper are vague and general, and the list is arbitrary and deficient. No fully objective, data-driven evidence base is given to support why these particular "harms" merit regulatory attention - or indeed why those explicitly excluded do not. It is notoriously difficult to establish a relationship between harm and content. Even where it seems intuitively obvious, the link may not be established in evidence, and it cannot be simply assumed that unpleasant content is harmful as this is often far from the case. The government must be careful not to allow emotion and distaste to drive this policy intervention, rather than evidence, as this risks regulation being disproportionate and even counter-productive. Additionally, if the standard for establishing risk is made easy to reach, in order to make it easier to compel action, then the risk of disproportionate action and over-censorship increase.

4.7 If a harms model is to be used, the harm to free expression caused by removing content or voices must itself be recognised and given weight in applying any duty of care. Any “harm-reduction” process must include consultation with experts in free expression rights, to ensure adequate protection.

4.8 Ultimately, many problems surrounding online content and social media do not relate to “harm”, but to the enforcement of limits that contracts impose on users. A harms-based regulatory model only addresses one slice of content, and does not acknowledge that even in the absence of harm Internet users have rights that need upholding and enforcing. To be successful, regulation needs to target company, rather than user, activity. It should address process-driven questions around how and to what extent platform architecture, content moderation and user redress mechanisms fulfil human rights obligations, including protecting democracy.

4.9 More holistic regulation in the online sphere is also needed to address the central issue of data exploitation by online companies - including data collection and retention, opaque advert targeting and recommendation systems and other algorithms. Upholding and enforcing principles of consent and fair processing would have a positive knock-on effect on user experience online, which will alleviate many of the government’s concerns driving this regulation.[9] Such regulation should, amongst other things, monitor whether users can control how their data is used and the kind of targeting and profiling that is employed. We note that this can be challenging to implement where regulatory scrutiny begins to intrude into legitimate commercial decisions.

5. A CALL FOR CO-REGULATION

5.1 The White Paper’s regulatory proposals would create an extensive model of state regulation, potentially leading to the censorship of millions of British citizens. This is unacceptable. State regulation has been deemed inappropriate for the press; it is hard to see how it could be justified here. Additionally, the real driver of content that is allowed or disallowed in online spaces is more typically driven by social and consumer expectations, rather than risk. This is not the business of the state. Direct government regulation, including through executive drafting of codes of practice (as proposed for terrorist content and child sexual

[9] Consider report by Bits of Freedom, *Fix the System, Not the Symptoms*, 19June 2019 <https://www.bitsoffreedom.nl/wp-content/uploads/2019/06/20190619-fix-the-system.pdf>

abuse material), or the creation of a government-but-independent regulator carries risks of being perceived as 'government control' of online speech, and of opening a process whereby further legislative requirements for restrictions on legal speech are added over time.

5.2 The government appears to have rejected independent self-regulation as a potential model. Independent self-regulation has the advantage of being independent from government; however, there may still be negative impacts on free expression.

5.3 We suggest a model of co-regulation. This offers a means to advance meaningful accountability and procedural improvements at company level, which would better protect human rights both where content is wrongfully removed and when it remains in place. Co-regulation recognises that laws are best enforced by governments. It creates public and parliamentary accountability for the kind of regulation that takes place whilst maintaining distance from state interference (including eliminating the potential of parliamentary pressure to take action in particular arising cases) and the setting of inappropriately restrictive norms.

5.4 Genuine co-regulation requires that the regulatory body be robust, independently managed, financially and substantively independent from both government and industry and with the power to make decisions that are final and respected. The precise nature of the relationship between state and private actors will determine the effectiveness of the co-regulatory framework: a range of stakeholders should be consulted on the design. A statutory footing and an audit mechanism is required for all to have confidence that the scheme is effective and accountable.

5.5 The impacts of the regulatory body on future press regulation needs to be considered. There appears to be some expectation that 'traditional' news media will be carved out of the Online Harms regulation; however, in that instance a state regulator for citizen speech content will stand in contrast to the non-state regulated speech of powerful publishers. It is hard to see the argument for allowing potentially racially-charged, provocative, scandalous, sexist or homophobic content and commentary on 'press' websites while the same content is methodically removed on social media. It is likely that media formats and social media will evolve together (lines are already blurring), making regulatory distinctions even harder to draw. In the short term, it seems impossible to expect newspaper content to be regulated differently when posted to social media: the White Paper's regulation will surely regulate newspaper content when shared on e.g. Twitter and Facebook, not least because tabloid articles are a classic vehicle for users to provoke the kinds of inflammatory commentary that the proposal seeks to limit.

6. REGULATION SHOULD BE EVIDENCE-BASED AND PROCESS-DRIVEN

6.1 Any policy intervention must be underpinned with a clear, objective evidence base which demonstrates that actions are necessary and proportionate. Regulation impacting on citizen's free speech needs to be based on evidence of impact traceable to specific pieces or types of content, activity or behaviour, rather than expectations or social judgements that these may be related to possible impacts. We note that it will be challenging to develop a regulatory scheme that fulfils this criteria.

6.2 We also note that even if content can be proven to be harmful, harm-reduction interventions need to be effective, rather than, for instance, driving content and behaviour into further unregulated spaces (e.g. the dark web) where there may be even greater potential for harmful effects. Evidence of necessity is essential to intervene, and removing content is not always the most effective means to mitigate.

6.3 Regulation should focus on reviewing internal company processes and auditing decisions. Given the enormous quantities of online content, we would expect powers of audit to be extremely important to produce robust results. This includes ensuring that platforms conduct sound content moderation with an effective appeals process to rectify mistakes.

7. NEED FOR COUNTER-NOTICE PROCEDURE IN CASES OF ALLEGED ILLEGALITY

7.1 The White Paper focuses on increasing responsibility for illegal or unwanted content; however, discussion around online content has to date given little emphasis to the listing of items for sale on online marketplace sites such as eBay or Amazon. Regulation of online content as the government proposes has a critical interaction with these types of sale listings. In this circumstance, strict application of content removal procedures by large platforms is infringing on free expression and being used by multinational rights-holders in an oppressive and anti-competitive manner. We propose that regulation require platforms to institute counter-notice procedures that enable users to take legal responsibility for their own content.

7.2 Since the E-Commerce Directive 2000 made platforms strictly liable for illegal content where notified of its presence [2.6], marketplace platforms such as eBay and Amazon have taken to using strictly applied takedown notice procedures to shield themselves from risk. If rights-holders e.g. patent-holders notify eBay or Amazon that a sale listing is for a product that allegedly infringes their legal rights, eBay and Amazon will immediately act to take down that listing and leave the issue to be resolved between the parties, which, for economic reasons, will rarely happen. The patent holder is thereby effectively granted absolute protection without proving their rights or the realistic possibility of legal challenge through a kind of privatised enforcement, whilst affected sellers have no recourse against the removal of their sales listings from the relevant platform.[10]

7.3 Multinational manufacturers are increasingly seeking plausible patents simply to spoil competition, knowing that smaller businesses cannot afford to raise revocation proceedings. Using these plausible patents as a blunt enforcement tool is known as "patent trolling". The patent holder is able to rely on the pure existence of their patent, and the alleged infringer has no means to present a counterpoint or rebuttal unless they can afford to challenge the patent itself (or in rare cases show that their compatible product does not infringe it).

7.4 eBay and Amazon's takedown-notice procedures are being exploited particularly by Epson, the multi-million-pound printer ink-cartridge manufacturer, to attack small online sellers. However, takedown exploitation is not a situation confined to the Epson experience. Material that relies on fair dealing in copyright law, such as quotation or parody exceptions in the UK, can be removed by a claimant with impunity in the same way. The originator cannot take legal responsibility for their content, therefore the platform operator must decide whether they want to take the risk from a copyright claim. The obvious response is to take down the content and remove the risk.

7.5 Many UK businesses, especially individuals and small traders, rely on Internet platforms to reach customers and conduct operations. The takedown-notice systems used by eBay and Amazon are applied inflexibly and give affected sellers no opportunity to challenge the removal of their listings and assert their legal right to post content. Yet it is the legal position as Amazon and eBay have no means under UK law to allow users to take legal responsibility for their actions.

7.6 In situations of online defamation, counter-notice is available as a method of challenging speech removal. It allows individuals to put up a direct defence to the removal of their post. Although an imperfect system, this style of rebuttal opportunity could apply well in commercial contexts to protect business

[10] See n.2 above at [7].

interests and ensure free speech rights. Unlike in cases of defamation, in this situation both sides have an interest in resolving the issue.

7.7 Under a counter-notice system, if the owner of the listing issues a counter-notice to takedown and the issuer of the takedown notice has no intention to go to court to assert their rights, then the listing will remain, protecting legitimate sellers. If on receipt of a takedown notice, the owner of the listing has no intention to resist, on the basis that they are outside UK jurisdiction or are aware that their listing is (or probably is) not legal, then the listing could be removed, upholding the law. The choice to counter-notice would however remain with the seller, granting them agency to decide how to proceed and protect (or not) their legitimate business and free- speech interests.

7.8 The counter-notice system might also assist in other illegal/unlawful content situations, such as alleged hate speech, where content may be removed mistakenly, or too swiftly. Platforms should be held liable where unambiguously illegal material is not swiftly removed on notification from a lawful authority, such as a court. However, until that point, if a user is willing and able to assert legal responsibility for their content, in principle they should be able to publish pending legal action.

7.9 Similarly with bullying. Bullying comprises a wide range of behaviour, from micro-aggressively "liking" posts, making consistently nasty comments to trolling, posting pictures without consent and posting information that can identify an individual's location (risking an offline attack) to targeted abuse and threats. Children receiving unwanted sexual attention from adults is also a form of cyberbullying. This is a complex issue which regulation cannot solve. A notice/counter-notice system could however operate on a privacy basis for users to request takedown of material about them. Additionally, in line with privacy protections, a takedown request procedure should enable swift removal of asserted non-consensual sexual images.

7.10 Individuals and organisations should also be able to access such removed content on request or assert a legitimate need for individual content to remain online by evidencing academic or journalistic necessity. Platforms should operate an appeals process capable of checking and rectifying removal errors, which are known to occur. Taking a graduating approach, content which is marked as being terrorist content or child sexual abuse material, for example, might only be put back after a review or appeal, rather than on a simple counter-notice.

8. SUPPORTING DEMOCRACY AND CRIMINAL JUSTICE

8.1 The White Paper gives little consideration to the interaction between social media platforms and democratic institutions, yet this is something that could positively be addressed through regulation. The Internet plays a significant role in enabling democracy by lowering barriers to participating in democractic discussion. As the barriers to participation lower, however, it becomes easier for vested interests to corrupt or distort these civic spaces. Combined with attention-based economics, this has led to an antagonistic online atmosphere punctuated by propaganda and "fake news", echo chambers and opaque political advertising. Regulation can play a role in diminishing this - although we suggest that data protection and electoral regulation are more likely to be effective in enabling free and fair political engagement than content-based approaches, especially since political speech is subject to high levels of protection in international law. Furthermore, online strife has 'offline' roots including wider social discontent. Attempting to improve online discourse without equally attempting to address social discord is unlikely to be very effective.

8.2 Regulation aimed at supporting democracy in online spaces could aim to address architectural issues around platform operation, such as how data is used and algorithms which manipulate the information environment through e.g. creating filter bubbles or promoting deliberately false content. Transparency over electoral advertising, spending and targeting are legitimate regulatory goals, and bodies such as the Electoral Commission and Information Commissioner's Office have relevance here. Platforms can also play a key role in identifying large-scale attempts at democratic interference by domestic or foreign powers.

8.3 Interaction between online and offline spaces is critical when it comes to illegal content. It is important for police action and prosecution to follow where criminal activity is suspected/indicated. Trust in regulation is built by there being real-world consequences for unlawful activity. Regulation should ensure that platforms have robust mechanisms and processes in place to prevent, identify and report to the police illegal activity. Children and young people in particular need easy access to mechanisms that allow them to alert platforms to potential offending including sexual exploitation and grooming. However, the privacy implications here are notable, and the regulator will need to take this into consideration when determining the point at which evidence should be handed to police.

8.4 Automatic removal should only take place where it is a copy of content already identified as unambiguously illegal, regardless of context. In all other cases, automation could be used to flag suspected illegal/harmful content, but the decision to remove should always be made by a human.

9. LAWFUL CONTENT IS LAWFUL CONTENT

9.1 It is clear that the government's starting point is that online platforms are hosting content and tolerating behaviour that whilst not illegal is unwanted. This is not a sound basis for regulation. Lawful content remains lawful, wherever it is disseminated. If the government wishes to ban certain types of content, it should legislate to make these illegal. Any such legislative initiative should be subject to proper consultation and parliamentary scrutiny, in order to ensure that any new laws are the result of a considered, rational process. Regulation generally needs to be complemented by other law reform around unlawful offensive communications: we refer to the Law Commission's work in this area.

9.2 The White Paper's category of "harms without a clear definition" (meaning lawful-but-harmful content) is unhelpful. This category exists nowhere else in law and the types of content included in this category are arbitrary and deficient. This categorisation has no place in a robust regulatory regime. Any bright lines should be between (a) clearly illegal content, (b) content that is not illegal but is nonetheless outside platforms' terms and conditions, and (c) content that is permitted under terms and conditions.

9.3 Platforms already have some incentives to attempt to balance free speech rights against questions around behavioural norms; in particular, there are reputational risks to overreaction in various directions, both in over-censoring and in permitting unpleasant material and activity to persist. The many kinds of unwanted content that platforms may be pushed towards banning also often reflect parts of human nature that are very hard to ban or regulate. Lines between open discussion, community support and promotion of harmful behaviour can further be hard to define.

9.4 Where regulation touches on lawful content, it should ensure that companies have appropriate content policies in place and adhere to these. Regulation should encourage companies to align their content policies / terms of service more closely with human rights law. It should also monitor whether platforms review reported content promptly and remove material and accounts - including bot accounts - that violate its terms and conditions. It should ensure that unwritten policies such as "newsworthiness" should play no role in this decision-making. Users should be given reasons where lawful content is removed, and have access to an effective independent appeals mechanism to challenge wrongful takedowns.

9.5 Regulation involving content take-down needs to consider at least four perspectives: the *complainant* challenging the content, which may be a state agency, corporate body or private person; the *platform* or *host*, which has enabled the content to be published; the *poster*, who may also have authored the content; and the *viewer*, who may have a right to access the content. At present, the White Paper focuses almost exclusively on the complainant and the removal of content to which they object, to the detriment of these other legitimate perspectives.

9.6 We recognise that offensive speech not reaching the bounds of illegal hate, and content aggregation in a way that is not feasibly possible in the offline world, can cause emotional distress and a toxic online environment for certain categories of platform user. Hower, this content may be protected by free speech rights. Hate speech is a controversial concept and remains without clear legal definition in international law. Thresholds of harassment may not be met if discrete postings are all legal and originate independently from multiple users. Platforms should have robust terms and conditions which make clear that hate speech and harassment violates their community standards. However, these also need to be clear on what does or does not constitute hate speech/harassment. This is necessarily a difficult balance to strike.

9.7 We welcome the call in the White Paper to "promote a culture of continuous improvement among companies and encourage them to develop and share new technological solutions rather than complying with minimum requirements." [intro, 14] It is important that regulation does not entrench monopoly positions but supports diversity in the online ecosystem. This support of diversity, however, is different to the White Paper's "duty of innovation" [Box 26].

10. CONSIDER THE FREE EXPRESSION IMPACT OF LIABILITY AND ENFORCEMENT

10.1 The White Paper's risk-based model of regulation focuses on companies removing "harmful" content. In this circumstance, any enforcement measure against non-compliant companies must be light-touch, as liability carries serious free expression risks.

10.2 It seems a peculiar policy goal to incentivise removal of legal content. Platforms typically disallow a range of legal content under their community standards, and it risks dangerous overreach for regulation to incentivise companies to go beyond what is already prohibited. If the duty of care success metric is e.g. decreased prevalence of certain types of specified content, there is a serious risk to legitimate expression, as this requires or incentivises wide-sweeping removal of lawful and/or innocuous content. We strongly caution against making companies liable for third-party content, as they will prefer elimination of legal risk and so be likely to over-correct in content removal and under-correct in user appeals.

10.3 Similarly, whilst we understand the need to remove some content speedily, e.g. live-streaming of criminal acts, excessive requirements in this respect pose risks to fairness and due process. Imposing time limits for content removal, heavy sanctions including personal liability for non-compliance [6.5] or incentives to use automated content moderation processes also heighten these risks.

10.4 Any liability and enforcement should focus on procedural requirements - asking, are effective policies and mechanisms in place? This being the case, it is hard to see how any enforcement measures for non-compliance beyond fines could be proportionate.

10.5 The White Paper considers relying on blocking powers [6.5] but these are incredibly blunt, a flawed technical solution to a complex societal problem. It must be remembered that a site that fails to comply with a regulatory regime is not the same as one whose users are consistently breaking the law. A block inevitably means the restriction of some or much legal content. Blocks always need to be limited to the very worst cases, where there is evidence to justify the request, and judicially authorised, not imposed by a regulator through administrative orders. Any request by a regulator to block a site must show respect for proportionality and the rights of the blocked site and its users. This might include the right for the site owner to be notified and to be able to stop a block in advance.

10.6 Our research into the state of copyright blocks has shown that even lists of legally mandated blocks of copyright material are ill-maintained. Injunctions against ISPs to block web services allows rights-holders to change the list of what is blocked as services change domains to evade the blocking orders. Over 30% of blocks we detected and examined in 2018 were incorrect, as no copyright infringing material was on those domains which were mostly unused and listed for resale. Most of these errors are still in place. While the impact of these blocks was commercial rather than speech, it is very worrying that organisations including the BPI and MPA are not taking their legal responsibilities seriously enough to ensure their blocking lists are kept up to date and accurate. Nevertheless, it also shows the challenge of the problem. Around 128 services blocked has resulted in around 3,000 domains being blocked, of which we have examined around 1,000. Maintaining and checking this number of websites, to ensure that they are in use and continue to be infringing, is presumably time consuming and costly, which is why it is being neglected. Nevertheless, it is important from a commercial and proportionality perspective that copyright owners continue to make the choice whether it is worth implementing and maintaining these blocks, rather than transferring this open-ended cost onto the public purse.[11]

10.7 Our research into adult content filters operated by UK Internet Service Providers (ISPs)[12] has identified high error rates, including wrongful blocking of legitimate advice sites, LGTBQ+ sites and even wedding services. We have also found cases where sites containing inappropriate content have not been blocked. There is no evidence that filters are preventing children from seeing adult content or keeping them safe online. The system of ISP filters is inherently problematic, as private companies are making questionable choices about what is and is not acceptable for under 18s, with no oversight or consideration of actual harms to young people.

10.8 Although the UK has used adult content filters since 2011, as a policy choice, ISP filtering is in fact prohibited under the EU Open Internet Regulation 2015, which requires all traffic to be treated equally, "without discrimination, restriction or interference".[13] Only where a legal process compels a block should ISPs restrict content, whether or not the user chose the restriction.

10.9 Technical capability of ISP blocking will be impacted by the DNS-over-HTTPS technical standards that is being adopted by the IETF. Blocking depend on disrupting the Domain Name Service (DNS) lookup process. DNS encryption methods will frustrate certain methods of blocking sites; we hope that this will in time reduce reliance on blocking as an enforcement mechanism, leading to alternative solutions and more nuanced approaches.[14]

11. TRANSPARENCY AND ACCOUNTABILITY

11.1 Regulation should be primarily and predominantly aimed at radically improving transparency and accountability on the part of social media platforms and others involved in the moderation and removal of online content. It should provide proper independent oversight of platform decision-making. Transparency should be a means to greater accountability, not an end in itself.

11.2 Transparency reporting must go beyond raw numbers and statistics. These can be problematic where requirements to publish data on e.g. quantity of takedowns or prevalence of certain types of content create perverse incentives for companies to simply take down borderline material, rather than grapple with

[11] See <https://www.blocked.org.uk/legal-blocks/errors> and <https://www.blocked.org.uk/legal-blocks>

[12] See n.3 above.

[13] Regulation (EU) 2015/2120, 25 November 2015, at [8] <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32015R2120&from=EN>

[14] For more on this, see our report, n.4 above.

hard cases and uphold free speech in pressured situations. Numerical data can also be blunt and unhelpful if presented without context. For example, if takedown rates increase, without context it would not be known if that was a positive result (more content was correctly identified for removal, terms and conditions being applied more effectively), or a negative result (more illegal content appearing on platforms, wider takedown approaches applied in violation of free speech). Similarly, if appeals mechanisms are being used more frequently, that can indicate users accessing due process mechanisms, but can also indicate a higher proportion of wrongful decisions that require correction.

11.3 Transparency also needs to be tailored for the regulator and for the public. Operational-level transparency presented for a general audience should qualitatively cover how and on what basis rules and policies are made, what factors inform content-related decisions and provision of hypothetical case examples showing how rules are interpreted and applied across a range of scenarios. It needs to include information around political advertising, with sufficient public information provided so that relevant third-parties can be held accountable.

11.4 Annual reporting is likely to lack utility, given the pace of social media and technological progress; any yearly progress report should be supplemented with more frequent reporting as part of an iterative improvement process.

11.5 Transparency also goes beyond reporting. Independent audits are essential for effective regulation. Audits go beyond ensuring moderation decisions are accurate and that inaccurate decisions and trends of decision-making are detected and resolved. They are needed throughout all parts of company systems, as the questions are about volume and impact of systems as much as particular aspects and decisions. This implies a familiarity with commercially sensitive information, so would be potentially more effective in a co-regulatory framework where technical expertise can be resourced. In any case, auditing will need to include means to examine algorithmic processing and machine learning techniques, which may be both controversial, for commercial reasons, and technically challenging for the regulator.

11.6 Reporting and accountability needs to address the impact on free expression that company decisions are having. Free expression is impacted by content removal, but also when people are unable to access speech arenas due to abuse. Impact can be people feeling too hurt or afraid to be online, or self-censoring out of an expectation of abuse that is both horrible and tiresome to deal with. Women and individuals from LGBT+ and black and minority ethnic groups receive a disproportionate volume and degree of abuse online. There is also a wider detrimental impact on society when people feel that views or feelings cannot be voiced. This risks causing further discontent and fuelling fringe views through legitimisation. It also risks suppressing important facts and opinions which are not widely felt to be acceptable, as ‘disturbing’ or ‘offensive’ tends to be used as a proxy for harmful. Any regulation system must contain mitigations to ensure that content is not over-censored. Platforms should be required to monitor and report on content removal on a regular basis to identify volume and categories of content being removed, overall free speech impact and any trends. As much as possible, boundaries of these kinds should be set by law.

11.7 Regulatory standard-setting for content moderation should be guided by the Santa Clara Principles.[15] Accountability includes developing quality standards for training content moderators.

11.8 Platforms should also be required to provide user-accessible information about the policies they have in place to respond to unlawful and harmful content, how those policies are implemented, reviewed and updated to respond to evolving situations and norms, and what company or industry-wide steps they have or are planning to improve these processes.

11.9 Content removal must be subject to precise, accessible and consistently-applied rules. Users must have effective ability to contest decisions made to remove or not remove content with appeals heard by an independent human decision-maker. A right to an effective appeal is essential for companies to fulfil human rights obligations.

11.10 If external actors are able to complain and remove material in bulk, there should be penalties for unjustified threats.

11.11 Algorithms and automated decision-making should not be developed or used in a way which would risk adverse impacts upon users’ human rights (such as the right to non-discrimination). There should be greater transparency over the use of algorithms, not least so that users have a basic understanding of when they are used and what their effects are.

11.12 The “super-complaints” mechanism referred to in the White Paper is undefined. It is not clear what the government’s thinking is in relation to this, and the regulator cannot act as a mass content-clean-up operation; however, we would welcome an equivalent to the provisions for the representation of data subjects under Article 80(2) GDPR, giving representative organisations a formal role in bringing fundamental-rights-based complaints.

[15] The Santa Clara Principles on Transparency and Accountability in Content Moderation <https://santaclaraprinciples.org>

11.13 In parallel, the government should enact the powers in Article 80(2) so that data protection complaints can be made on behalf of people who are unable to easily identify their data rights. Many of the underlying problems of social media relate to abuse of personal data, thus improving data protection enforcement can be a positive force for content - see 2.5 above. Content regulation and privacy regulation can be seen as complementary: in many cases, privacy and personal data is *input* or *cause* and content regulation focuses on *output* or *effect*. It is often wiser to regulate to deal with cause rather than effect. In this case, data protection concepts such as 'fair processing', 'consent' and 'sensitive personal data' are very powerful levers to ensure that algorithmic assessments by companies can be made accountable to individuals and wider social concerns.

12. ACCOUNTABILITY FOR STATE ACTORS

12.1 Whilst company transparency is an important focus in regulation, it is equally important that State actors removing content should be accountable through being subject to independent authorisation and supervision. Takedown requests made to platforms under terms and conditions from government bodies including Police Intellectual Property Crime Unit (PIPCU) of the City of London Police, the Counter-Terrorism Internet Referrals Unit (CTIRU) of the Metropolitan Police and others must be included in any transparency reporting. The work of Nominet and the Internet Watch Foundation (IWF) should also be considered. Government should set out a legal framework that includes prior authorisation for content removals by authorities, organisations such as the IWF and for external accountability.

12.2 As noted above at 2.9, we are concerned that platforms will tend to continue to rely on terms and conditions for takedowns, and that regulation likewise will push towards the way that terms and conditions are enforced, through detection and removal of content, and thus towards removal of legal content, including content that does not breach terms and conditions.

12.3 At present, CTIRU and PIPCU among others rely on breach of terms and conditions as a reason for platforms to remove content or for Nominet to suspend domains they consider to be unlawful. Yet there is no independent assessment of the lawfulness of the sites or content in question before notifications are made.

12.4 In CTIRU's case, this has led to some clear mistakes in notifications, for which they are not accountable.[16] Indeed, platforms themselves say they do not remove all content requested by CTIRU, pointing to the likelihood that CTIRU does not always accurately identify material which is unlawful and wrongly considers some lawful material to be unlawful. Users are not necessarily

[16] See n.1 above for more detail on this and below CTIRU/PIPCU/Nominet/IWF/BBFC points.

informed that their actions are potentially unlawful, nor that content they have attempted to view is potentially unlawful. There is no opportunity for users to seek redress in cases of wrongful removal. As the same content may be removed on multiple platforms, it is important that CTIRU can be asked to stop removal requests when they make a mistake.

12.5 CTIRU has consistently resisted requests for information to establish more information about their work. It seems clear that Freedom of Information requests have been routinely denied without any real assessment: our own request to the Metropolitan Police for a list of statistics held was denied under 'national security' grounds, only for an appeal to the ICO to find that the information did not exist. In any case, the lack of routinely held statistics is worrying as it does not indicate good performance management. The sensitivity of CTIRU's work in our view normally falls well below that of normal policing. It is clear that someone is 'watching' groups and material through takedown, a kind of 'tipping off'; and destruction of evidence is the natural result of a takedown request. Thus in our view blanket claims of national security are unreasonable in relation to CTIRU's work as a de facto censor. Instead, CTIRU should be independent of the police, and work to be transparent and accountable to the public.

12.6 In the case of the Internet Watch Foundation, appeals can be made and are made successfully, despite the fact that assessments of child abuse images ought to be a clear-cut matter. This should remind us that removals of extremist content and domains for criminal activity at volume will also inevitably feature mistakes.

12.7 Block pages are not required by the IWF, so people accessing certain material are not always warned that the content is likely to be illegal. The IWF's blocking regime is not consistent with the EU Open Internet Regulation, which envisages that Internet blocks should be put in place by ISPs only as the result of a legal process. An assessment of legality by the IWF is not sufficient. If it were, it would be open for ISPs and various private actors to block a large amount of content of varying types, ranging from defamation to copyright infringement, without recourse to a court. An independent authorisation process would reconcile IWF practice with current law.[17]

12.8 PIPCU and over ten other authorities notify Nominet of over 30,000 domains a year to be suspended for reasons ranging from the sale of counterfeit watches and handbags through to sale of unlicensed pharmaceuticals.[18] It is unclear if an independent appeal system currently exists, although Nominet state they will

[17] See 10.8 and n.13 above.

[18] See <https://wiki.openrightsgroup.org/wiki/Nominet/Domain_suspension_statistics> for suspension statstics and responses to requests for policy information.

put one in place. In any case, there is no prior independent assessment of the domains PIPCU and others ask to be removed. Many of the authorities making requests do not have policies or do not publish policies explaining when they may ask Nominet for domains to be suspended. Splash pages explaining why domains are suspended by Nominet and law enforcement are absent, which could lead to consumer harm, if people are not advised that materials they have received could be dangerous.

12.9 We are also worried by the potential for administrative blocking of websites by the BBFC to extend over time to many thousands of websites that are publishing legal adult content. This may impact sexual minorities such as LGBTQ+ disproportionately while making little impact on child safety. Website blocking should require independent authorisation, so that questions of proportionality can be properly assessed, rather than being implemented by a regulator (see also 10.5 above).

12.10 There is no consistent process or standard for content removed at the request of law enforcement. Sometimes appeals exist, and at other times they do not. There is no prior authorisation and it is unclear how or if oversight bodies check the work of these bodies. Accountability should be clarified and prior authorisation, appeals and notification processes created as a result of the White Paper response.

13. USE REGULATION TO BUILD PUBLIC TRUST

13.1 Platforms are private companies and operate differently according to internal company identity and policy. The diversity in the platform ecosystem is positive and support innovation. Nonetheless, consistency in compliance with fundamental rights and transparency across platforms would increase public trust. Regulation should increase public trust that online terms and conditions are a genuine two-sided contract and will be adhered to and enforced.

13.2 It would increase trust to demonstrate that the government is acting in the public interest by protecting children and vulnerable groups in a way that upholds and protects their fundamental rights, including their rights to freedom of expression.

13.3 It is important for police action and prosecution to follow where criminal activity is suspected/indicated. Trust in regulation is built by there being real-world consequences for unlawful activity.

13.4 As detailed above, regulation should focus on systemic issues. Separately, an independent dispute resolution mechanism should be established to facilitate mediated conflict resolution between platform users. This could improve individual access to effective remedy in appropriate cases without overburdening the courts and support improved civil discourse on platforms.