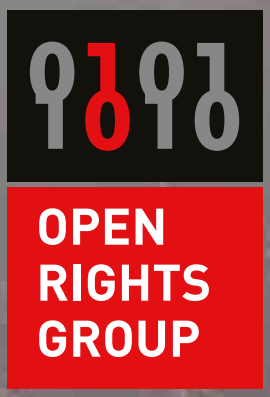


UK INTERNET REGULATION PART II

About Jim Killock, Executive Director

Since joining Open Rights Group in January 2009, Jim has led campaigns against three strikes and the Digital Economy Act, the company Phorm and its plans to snoop on UK users, and against pervasive government Internet surveillance. He is working on data protection and privacy issues, as well as helping ORG to grow in size and breadth. He was named as one of the 50 most influential people on IP issues by Managing IP in 2012. In the same year ORG won Liberty's Human Rights Campaigner of the Year award alongside 38 Degrees, for work on issues from copyright to the Snooper's Charter.

Since 2009, ORG has doubled its supporter base, budget and workload, and held five activist conferences, ORGCon.

Jim is a trustee of FreeUKGen, the volunteer project to digitise genealogical records, and sits on the Governance Board of CREATe, the UK's research centre for copyright and new business models in the creative economy. He is on the Advisory Council of the Foundation of Information Policy Research.

Before joining ORG, Jim worked as External Communications Co-ordinator of the Green Party. At the Green Party, he promoted campaigns on open source, intellectual property, digital rights and campaigned against the arms and espionage technologist Lockheed Martin's bid for the UK Census. Lockheed Martin have since been prevented from handling UK Census data as part of their contract. He was also a leading figure in the campaign to elect their first party leader, Caroline Lucas MP. He has a blog at http://jim.killock.org.uk/

PGP key available via: pgp.mit.edu

OPEN RIGHTS GROUP

About Open Rights Group

As society goes digital we wish to preserve its openness. We want a society built on laws, free from disproportionate, unaccountable surveillance and censorship. We want a society in which information flows more freely. We want a state that is transparent and accountable, where the public's rights are acknowledged and upheld.

We want a world where we each control the data our digital lives create, deciding who can use it and how. We want the public to fully understand their digital rights, and be equipped to be creative and free individuals. We stand for fit-for-purpose digital copyright regimes that promote free expression and diverse participation in culture.

We campaign, lobby, talk to the media, go to court — whatever it takes to build and support a movement for freedom in the digital age. We believe in coalition, and work with partners across the political spectrum.

We uphold human rights like free expression and privacy. We condemn and work against repressive laws or systems that deny people these rights. We scrutinise and critique the policies and actions of governments, companies, and other groups as they relate to the Internet. We warn the public when policies — even well-intentioned ones — stand to undermine the freedom to use the Internet to make a better society.

www.openrightsgroup.org

EXECUTIVE SUMMARY

Independent, accountable and transparent decisions

1

This report follows our research into current Internet content regulation efforts, which found a lack of accountable, balanced and independent procedures governing content removal, both formally and informally by the state.

There is a legacy of Internet regulation in the UK that does not comply with due process, fairness and fundamental rights requirements. This includes: bulk domain suspensions by Nominet at police request without prior authorisation; the lack of an independent legal authorisation process for Internet Watch Foundation (IWF) blocking at Internet Service Providers (ISPs) and in the future by the British Board of Film Classification (BBFC), as well as for Counter-Terrorism Internet Referral Unit (CTIRU) notifications to platforms of illegal content for takedown. These were detailed in our previous report.[1]

The UK government now proposes new controls on Internet content, claiming that it wants to ensure "the same rules online as offline". It says it wants "harmful" content removed, while respecting human rights and protecting free expression.

Yet proposals in the DCMS/Home Office White Paper on Online Harms[2] will create incentives for Internet platforms such as Google, Twitter and Facebook to remove content without legal processes. This is not "the same rules online as offline". It instead implies a privatisation of justice online, with the assumption that corporate policing must replace public justice for reasons of convenience. This goes against the advice of human rights standards that government has itself agreed to and against the advice of UN Special Rapporteurs.[3]

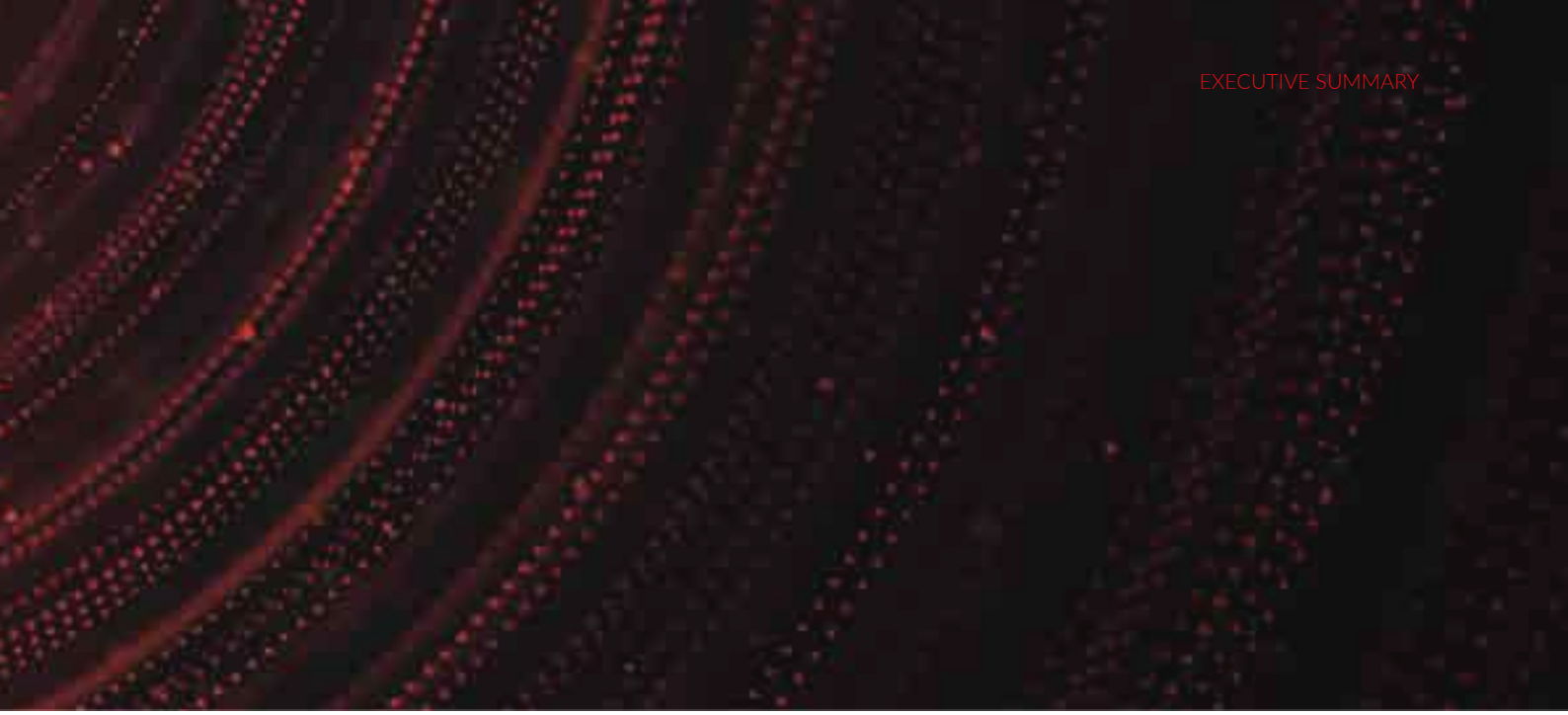
The government as yet has not proposed any means to define the "harms" it seeks to address, nor identified any objective evidence base to show what in fact needs to be addressed. It instead merely states that various harms exist in society. The harms it lists are often vague and general. The types of content specified may be harmful in certain circumstances, but even with an assumption that some content is genuinely harmful, there remains no attempt to show how any restriction on that content might work in law. Instead, it appears that platforms will be expected to remove swathes of legal-but-unwanted content, with as as-yet-unidentified regulator given a broad duty to decide if a risk of harm exists. Legal action would follow non-compliance by a platform. The result is the state proposing censorship and sanctions for actors publishing material that it is legal to publish.

Demands from the government point in contradictory directions. It wants social media platforms to protect free expression, but it also wants platforms to remove material that it deems morally offensive. To an extent, the drive towards content removal reflects concerns in society, and certainly the media. There is a large amount of hate, bigotry and prejudice online, which makes it easy to argue that some content is morally abhorrent and should not be allowed by platforms, even if it is not illegal. It is also easy to point at negative user conduct on platforms and conclude that because this is within a defined online space it is primarily the responsibility of the platform to stop and perhaps prevent such incidents, even if concrete proposals about how to do this seem likely to restrict legitimate free expression. However, this conveniently sidelines the principle that platforms

1 https/:www.openrightsgroup.org/about/reports/uk-internet-regulation

2 https://www.gov.uk/government/consultations/online-harms-white-paper

3 Ibid, p3-4



should – as the government says – apply the same rules for free expression online as offline. What is legal is legal.

Central to our concerns are that users have a right to publish legal material, which must be upheld. The EU legal framework, notably the E-Commerce Directive 2000, has been critical in securing the ability for platforms to operate without unfair risks. As a result, it has been portrayed as unduly restricting action against platforms. However, the regime in fact offers very little protection to platforms in the face of a well-formed notice as this constitutes "actual knowledge" of something potentially illegal and removes platforms' protection from liability.

The framework further offers no protection to users, who cannot by default defend their right to post legal content that has been made subject to a notice. In the case of Intellectual Property allegations, this has made platforms such as eBay very cautious, to the financial detriment of legitimate UK businesses. Copyright takedowns run on US procedures, which allow users to assert their right to publish, but only on the basis of accepting the jurisdiction of US courts. Only in libel law does a sufficient notice and counter notice system allow users to assert their right to publish.

We propose that a focus on transparency and ensuring that processes enforce terms and conditions are the right approach, including audit functions to assess overall performance. Regulation of platforms needs to be independent of both platforms and government to maintain long-term confidence. This could mean a form of co-regulation that allows government to define some of the objectives of regulation and ensure that it is independent of the companies that are regulated. Functions need to include audit of systems to ensure errors are detected and reduced.

Other measures to support a focus on process and transparency must include notice and counter-notice systems. These are vital for users to defend the legality of what they publish. The UK already has this model for defamation.

Measures could include introducing alternative dispute resolution (ADR) in specific circumstances to deal with specific kinds of private disputes about unlawful material. ADR could also deal with some disputes about behaviour on platforms that is potentially in breach of terms and conditions, although there are many limitations to this approach. In particular ADR is *not* appropriate as a means to enforce restrictions on speech that may be found within terms and conditions.

Proposals for automated content identification are often extremely error-prone, as we show in our 2019 Blocked report on Internet filters.[4] They should not be forced on platforms by law. Decisions on content removal should be dependent on human intervention where there is any likelihood of doubt. The principles of 'no general monitoring' and 'no pre-censorship' should be upheld.

4 https://www.openrightsgroup.org/about/reports/blocked-2019 ; see also https://www.blocked.org.uk/stats

INTRODUCTION 2

a. Purpose and scope of report

This report is written in response to a government White Paper which sets out goals for Internet content regulation in the UK. The government has made it clear that its primary concern is removing 'harmful' content from the major social media platforms (hereafter "platforms").[5] It is also considering legal changes to make platforms liable for content.[6]

The report addresses UK content regulation at online platforms: issues with current systems, mechanisms and frameworks, and recommendations for the future. We examine the online content liability regime and show that the current arrangements are weak, and fail to protect users from spurious takedown requests. We argue that liability protections are fundamental if users are to be able to use third-party platforms to publish legal material free from interference, and that the task for the government should be to enable all parties to seek redress, whether they want content removed or need to defend their right to publish.

We further consider other approaches to content regulation that have been advanced, including the idea of a 'duty of care' imposed on online platforms, which has in the White Paper been adopted by the government as its preferred way forward. We also look at alternative dispute resolution to show where it may be useful as well as its limitations and briefly discuss co-regulation.

b. Defining the problem

i. Right to redress, right to publish, right to defend

Content regulation 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.

Too often, policymakers addressing content regulation have considered only the positions of the *complainant* and the *platform* or *host*. Their general approach is that complaints about individual pieces of content need to be swiftly dealt with and are not being properly addressed. The idea is advanced that platforms are failing to find problematic content and remove it. The rights of *poster* and *viewer* to express themselves and access information are either diminished or sidelined. Furthermore, the contention is made that there is too much material to deal with, so processes must abandon notions of external accountability, fairness or balance of rights.[7]

ii. Evidence

A feature of many of the online content debates is a reliance on thin evidence, or no evidence, with various actors preferring to frame issues through emotive narratives. For example, recent copyright debates have featured exaggerated claims about damages and debates on pornography feature excessive and emotive claims about harms to children. Too often, it is emotion and distaste rather than evidence which appears to drive policy interventions, which risks them being disproportionate and even counter-productive.

---

5 https://www.gov.uk/government/news/new-laws-to-make-social-media-safer

6 https://www.gov.uk/government/consultations/internet-safety-strategy-green-paper See page 17 "Online platforms need to take responsibility for the content they host. They need to proactively tackle harmful behaviours and content. Progress has been made in removing illegal content, particularly terrorist material, but more needs to be done to reduce the amount of damaging content online, legal and illegal.
We are developing options for increasing the liability online platforms have for illegal content on their services. This includes examining how we can make existing frameworks and definitions work better, as well as what the liability regime should look like in the long-run.

7 Processes that platforms are expected to put in place do not respect the procedural safeguards that would be expected under Article 6 ECHR

It is often assumed that unpleasant content is harmful: this is taken as self-evident and requiring no further investigation. However, this is often far from the case. Many unpleasant kinds of content may be socially unacceptable, legal and have very limited potential for harm simultaneously. Harm may be suspected but in fact negligible.

Establishing that a general societal harm exists is easier than drilling down into what specific categories of speech content may cause specific incidences of harm to specific identified individuals or groups, and what specific mitigations might be appropriate to counteract the likelihood of this. The government has done some thinking around this - it has attempted to establish what content may be potentially harmful for children to encounter, for instance. However, as the UKCCIS Evidence Group Reports 2017 highlights, "it remains difficult except in retrospect to pinpoint the moment when children succumb to specific online risks" and it notes that children "already at risk offline are more likely to be at risk and vulnerable online."[8]

The report shows that risk is a complex area. Harm-reduction interventions need to be effective, rather than, for instance, driving behaviour and content into further unregulated spaces where it has greater potential for generating harmful effects. Risk mitigation should not rely solely or mostly on platforms but be centred around "developing critical ability and technical competency in terms of education, as well as supporting children online and offline through constructive and informed parenting practices, through safety and privacy by design, and by improving the digital expertise of relevant welfare and other professionals who work with children." Even where a specific harm is identified, therefore, mitigations aimed at content rather than users may need to be narrow and focused to be proportionate.

Developing effective policy relies on it being underpinned by objective, data-driven evidence. Government and parliament also need time to digest technical and expert information and understand the policy considerations. This will be much more acute after Brexit as the fundamentals of digital policy may be their direct responsibility. For this reason, we recommend that both institutions significantly increase their capacity to deal with the detail of evidence and policy.

Recommendation 1

Increase the capacity of Government to deal with Internet issues

Recommendation 2

Increase the capacity of Lords and Commons to conduct detailed research, scrutiny and policy work

iii. Lawful and unlawful content and behaviour

The debate that the government has initiated has great potential to blur distinctions between lawful and unlawful content, as well as criminal and civil wrongs. The state is entitled to intervene when there is a clear public danger posed by someone's actions, such as incitement to racial hatred, criminal levels of copyright infringement or criminal harassment. However, racism that is unpleasant but does not pose a direct threat, instances of possible copyright infringement that may be legitimate uses, or robust but rude arguments online do not necessarily pose such clear public danger. In order for the criminal law to be justifiably applied, harms must be demonstrable and sufficiently serious enough to mean that there is a public interest in the state intervening. This has to be the case in order for laws to command public support.

Criminal and civil law matters also need to be subject to due process. Accused persons need to be able

8 *Children's online activities, risks and safety A literature review* by the UKCCIS Evidence Group (2017) https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/650933/Literature_Review_Final_October_2017.pdf

to defend themselves. It cannot be for a platform to unilaterally determine whether action is legal or illegal without allowing the subject of their decision meaningful recourse to further review. Perpetrators of crimes should be tackled and this is best done through judicial processes.

Potential for blurring of policy objectives arises when Internet platforms are asked to take measures against unwanted content or behaviour. In these cases, policies may aim at giving platforms legal incentives or duties to act against lawful content, perhaps based on a notional or vague category of harm. This must be avoided. Instead, the government must identify actual categories of harm that are demonstrable and clear, if the law is to be used to restrict content.

iv. Rights or risks

As will be seen in our discussion of a 'duty of care' below, the government favours an approach based on the idea of risk posed to Internet users. This can be appropriate in the case of clearly criminal content and activity, for instance for spam or phishing. However, when risks are harder to discern, only apply to certain people, or are wider social risks rather than personal risks, the case for intervention becomes harder and the potential for overreach becomes greater. For most speech, where the questions often amount to civility rather than harm, risk is not an appropriate model. Focusing on harm naturally produces models of content removal rather than fair, necessary and accurate actions.

We favour a rights-based approach. This gives policymakers the ability to balance the needs of all users. It also focuses policy on process, which is the precursor of accuracy and balance.

v. Business models and incentives

It is commonly assumed that platforms allow unpleasant or harmful content because it is profitable. In our view, although this is an oversimplification, it is true that content may be prioritised by platforms or circulated by users because it is 'appealing'. Both users and platforms seek attention. For users, this may be about prestige, a sense of fun, or a desire to influence. For platforms, the motivation is that users want to spend more time using their product.

For example, content may have 'viral' qualities, which range from the amusing to the shocking. Platforms will often seek to ensure this interesting content reaches people more quickly than other content, which sets up the possibility that untruthful, exaggerated or emotive content may be more likely to 'succeed' in online spaces where more balanced or nuanced content will not: much as it does in other media contexts. When this competing attention is commercially driven, it also creates the incentive for media outlets to produce content that is as appealing as possible for as low cost as possible.

These are not new issues. They have been found in other media at other times where low production costs have dominated the market. For instance, cheap or free newspapers have suffered from a poor reputation compared with paid-for, subscription-based news services.[9] For the purpose of this policy debate, we should understand that interventions aimed at regulating particularly extreme content may be relatively limited in their impact if the underlying business model does not change.

9 See Holiday, Ryan (2012) *Trust me, I'm lying*, New York for instance; or Wu, Tim (2016) *The Attention Merchants: The Epic Scramble to Get Inside Our Heads.* New York.

INTERNET INTERMEDIARIES AND LIABILITY

3

Legal protections which shield platforms from intermediary liability for third-party content are the foundation on which online services rely to allow a wide range of individuals to use and publish their own works. They enable companies to be able to operate without the threat of immediate and unreasonable action and thereby also protect users' ability to exercise their right to freedom of expression. However, the shield is vulnerable. A properly formed notification of legal infringement by a rights-holder will swiftly remove liability protection; a fact which is often ignored in policy discussions.

In this section, we examine the deficiencies of the liability framework, which we believe are to be found in the lack of additional frameworks for takedown notices. In particular, this leaves users unable to defend their right to publish except in the case of libel claims. We also examine the possible proposals for change to liability protections that are being discussed as part of the government's online safety strategy.

a. Current framework

i. Protections for platforms publishing user content

For the purposes of the Digital Charter and online safety discussion, which is focused on the role of Internet platforms, the most important legal instrument is the European E-Commerce Directive (2000/31/EC).[10] This is widely said to protect intermediaries from liability - but this is untrue. In Europe, platforms that allow users to publish content are shielded from liability until they have actual knowledge of unlawful activity. In practice, this means receipt of a well-formed notice removes liability protection, leaving the platform in the same legal position as a traditional publisher.[11]

There is no possibility in the UK for the poster of content subject to a notice of illegality to challenge the issue of a notice or assert their legal position and maintain their right to publish. The exception is libel law in England and Wales, where users can assert their right to publish through issue of a counter-notice.

The E-Commerce framework was not intended to be the end of discussions about online content removal. It was, we believe, expected that processes would be developed that would be tailored to the needs of differing kinds of complaint. We welcome ongoing discussions at EU-level about its review.

The lack of a more general liability protection and of specific processes to place customer and compainant together already leads to significant cautiousness in some instances. eBay and Amazon in particular accede to all requests by intellectual property rights-holders

10 https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:32000L0031

11 For a fuller discussion see https://wiki.openrightsgroup.org/wiki/Intermediary_liability

to remove marketplace posts that appear to infringe their rights, with no meaningful recourse available to affected UK traders. Copyright material hosted in the USA is subject to a notice and counter-notice system, but UK businesses have no such equivalent opportunity. As matters stand, they are left at the mercy of platforms and cannot rely on notice-and-takedown procedures to act fairly to their customers.

ii. No General monitoring obligations

The E-Commerce Directive prohibits general monitoring by platforms of user-generated content. The rationale for this prohibition was to prevent the establishment of crude, automated content inspection and filtering techniques through national legislation; for instance, rights-holder organisations demanding that ISPs inspect Internet traffic for transmission of copyright works, in order to block it. The prohibition is a sensible precaution against disproportionate use of technology to establish suspicionless surveillance for a variety of possible behaviours; a tempting but unwise policy direction.

The protection against general monitoring however does not take direct effect in EU Member States and has not yet been transposed into UK law. It may be found to exist through case law or other mechanisms, however we recommend that it is made explicit through statute.

Recommendation 3

Incorporate the No General Monitoring obligation into UK law

iii. Case study: eBay, Amazon and UK cartridge resellers

eBay UK has interpreted the risks to their company from intellectual property infringements as substantial, as the protections they have as host are very weak in the EU. The result is that they recognise any rightholder's notice for takedown as valid, so long as they sign up for the "Verified Rights Owner" (VeRO) programme.[12] The programme allows any verified rights-holder to remove any listing from eBay, as long as they assert their rights to the original poster beforehand.

This programme has allowed Epson, the well-known ink printer cartridge manufacturer, in the UK to remove for-sale listings of compatible ink cartridges on the basis of a claimed patent infringement. These cartridges are predominantly sold by small UK enterprises, who operate mainly on marketplace platforms and sometimes additionally through their own websites. There is no system of counter-notice in the eBay system or under UK law, so companies have simply found their listings notified and then removed.

If the patent were ever to come to court it would be very controversial, as it attempts to claim that water-resistant connector arrangements cannot be copied without licence. This would open up a strategy for any manufacturer to control secondary markets through patenting connectors with certain useful properties – in essence being a patent on a plug.

Attempts to control secondary markets are always controversial and could easily be held to be against the public interest. The use of a patent to prevent the use of connectors would be controversial and could well be contentested. In this case, however, Epson has had no need to show that the patent is valid in court, as the VeRO programme has allowed it to take action without reference to judicial proceedings.

Amazon has taken a similar view in relation to listings on its open marketplace, and responds positively to Epson's requests for listings to be removed.

Epson's activities seem instinctively unfair, especially as it has not tried to take the cartridge importers to court.[13] When Epson acts to remove listings it has a vast amount more power than the small companies selling cartridges in this dispute. There is little the cartridge sellers can do to fight back, particularly as it seems there is no prospect that the validity of the patent itself will be tested in court. It would be very hard for the very small sellers to attempt this purely because of the financial risks to them, even if their arguments are sound.

The same problems around takedown will be evident for other eBay and Amazon resellers, in the fields of trademark and copyright. It would be useful for the Government or the Intellectual Property Office (IPO) to try to collect evidence of this, in order to see whether a notice and counter-notice regime would benefit UK businesses and website users.

Our recommendation is that a system of notice and counter-notice, backed by the option of alternative dispute resolution (ADR), is needed to reduce abusive and wrongful notifications. We also encourage the IPO to review how the patent system operates, to see whether there are means to limit what is effectively an abuse of a this legal right.

12 For more background and links to Epson's programme, see https://wiki.openrightsgroup.org/wiki/Epson

13 http://www.legislation.gov.uk/ukpga/2017/14/contents/enacted The Intellectual Property (Unjustified Threats) Act 2017

b. Potential changes to platform liability protections

i. Unwanted content at platforms

The government's starting point is that platforms are hosting content and tolerating behaviour that although not illegal is unwanted. For instance, DCMS stated to the House of Lords inquiry, the output report of which was published 9 March 2019, that its current priorities include:

> **"Online harms - protecting people from harmful content and behaviour, including building understanding and resilience, and working with industry to encourage the development of technological solutions.**
>
> **Liability – looking at the legal liability that online platforms have for the content shared on their sites, including considering how we could get more effective action through better use of the existing legal frameworks and definitions."**[14]

DCMS added that they wish to:

> **"harness the ingenuity of the tech sector, looking to them for answers to specific technological challenges, rather than Government dictating precise solutions … consider the full range of possible solutions, including legal changes where necessary, to establish standards and norms online."**

In more detail, DCMS states (our emphasis) that they want Internet companies to:

> **"proactively tackle harmful behaviours and content on their platforms. Progress has been made in removing illegal online content, particularly terrorist material and child sexual abuse and exploitation material, but more needs to be done to reduce the amount of damaging content online, both legal and illegal. As the Prime Minister announced in January 2018, we are looking at the legal liability that social media companies have for the content shared on their sites."**

This direction of policy is reflected in the White Paper. The approach suggested creates liabilities for platforms that allow third parties to behave in ways that are legal but deemed to be harmful (however that is defined), and probably incentivising the detection and removal of illegal content by platforms. The removal of legal material by government instruction is obviously problematic, as is removing material without human review. The UN Special Rapporteur on Freedom of Expression and Information has specifically warned against 'pre-censorship' of material.[15]

In the White Paper published in April 2019, the government seems to hope that technological solutions will play a significant role in eliminating unwanted content. No definition or evidence base for harm is established. Instead, it appears that the harms are assumed to be self-evident.

Platforms already have some incentives to attempt to balance speech freedom against questions of behavioural norms. There are reputational risks to overreaction in various directions, through claims of abuse of personal data, over-censoring or permitting unpleasant activity or content. Indeed, these issues may already creating a toll on Facebook's user base, and have economic impact on their share price.[16]

For companies making decisions about allowing or disallowing content, risks arise whichever way they turn. Removal of content and failure to moderate can both lose user base or damage reputation. When content restrictions are created by companies or by government policy, the result can be counter-intuitive. For instance, US laws against online promotion of sex work, on the grounds of its association with sex trafficking, led to more dangerous street working for some, and subsequently the creation of a non-US Internet platform called 'switter' to cater for sex workers banned from advertising on US platforms, which now has 125,000 users.[17] Thus apparently successful policies restricting content may in fact push users further out of reach, especially if they have a genuine (and in their view legitimate) wish to communicate with each other.

The many kinds of unwanted content that platforms may be pushed towards banning often reflect parts of human nature which are very hard to ban or regulate. These include prurience, enjoyment of offending or causing over-reaction, sexual lewdness and

---

14 http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/communications-committee/the-internet-to-regulate-or-not-to-regulate/written/86136.html

15 See chapter 2

16 *Over $118bn wiped off Facebook's market cap after growth shock* The Guardian 26 July 2018 https://www.theguardian.com/technology/2018/jul/26/facebook-market-cap-falls-109bn-dollars-after-growth-shock

17 https://switter.at

fascination with gore and death. The lines between self-help and promotion of harmful behaviour can often be hard to define.

Interventions therefore need to take account of the motivations of the actors and their likely response as well as whether it is appropriate for government to seek to legislate for action against content that is disturbing or unpleasant but ultimately not designated by their parliamentary leadership as illegal.

ii. Unwanted behaviour

A general assumption in law and policy is that there is a difference between free expression of views and a course of conduct involving communications that are aimed at a particular harm, such as harassment, bullying or libel. This makes a great deal of sense from a legal perspective, however it poses problems when considering the kinds of changes that Internet platforms might make.

Firstly, courses of behaviour are often harder to judge than content alone. This is because the motivation of the poster may need to be understood, rather than just the content.

Secondly, illegal and unlawful conduct is likely to be a much worse kind of behaviour than would be generally regarded as unacceptable. For instance, insulting someone may not be libellous, but is generally regarded as unacceptable. Persistent rebukes are unlikely to constitute harassment or intimidation, although they are socially unacceptable. Thus a platform or individual is likely to want to act on a wider range of behaviour than legal limits permit.

Thirdly, the gaming of systems is a natural consequences of creating powerful online tools. Accessibility and low costs in digital systems create a means for bad actors to try to find ways to abuse them for their own ends. Spam, viruses and fake content are well-known examples of digital gaming. Many kinds of socially abusive behaviour are similar in nature to gaming of systems, for instance when certain individuals target celebrities for abuse they are simply taking advantage of the platform's qualities that permit this to occur.

Another behaviour that is very hard to deal with is attempts to belittle or pressurise individuals, for instance on Twitter, where pile-on crowd effects can lead to very traumatic experiences for users who experience a high level of public opprobrium. When this happens spontaneously, it may be extremely damaging for an individual, especially if they have additional vulnerability. However, it also may be very hard to state with certainty that anyone in particular has acted criminally, or to conclude that Twitter has a direct responsibility or duty to make such behaviour stop. It is also far from clear that it should be the state's responsibility to direct Twitter as to when to take action and what action to take to police its own commercial space.

As noted above, platforms have commercial incentives to ensure their users are comfortable and find their experience on the platform fulfilling, and to limit the activities of bad actors. Whilst platforms may wish to tolerate a certain level of bad behaviour in order to maximise their user base, in the long term the presence of bad behaviour can lead to platforms losing business. A small percentage of disruptive users can create significant problems for platforms. The impact of this can be seen where newspapers have closed comments sections rather than expend effort overcoming the difficulties created by problematic individuals.[18]

It may be that network effects and the investment of other actors in their presence on platforms like Twitter and Facebook are militating against the desire of users to find better or alternative tools. Nevertheless, platforms have experienced commercial damage from the loss of users as the result of recent controversies such as the Cambridge Analytica scandal.[19]

iii. Incentives to remove content

Any incentive to remove content swiftly is not the same as an incentive to do this accurately. It is much easier for policy to concentrate on speed, which is much more easily measured than accuracy, which is subjective and hard to ascertain.

It is likely that policies aimed at making companies take action will come at the cost of legitimate expression and accuracy. Companies of all sizes will not wish to create new costs for themselves. Large platforms in particular which are most likely to be subject to scrutiny and enforcement will prefer automation to human review, and prefer elimination of legal risk.

Any content removal policy will be particularly problematic if systems of notice and counter-notice are not present, so that users can assume legal responsibility for their actions and defend their right to publish.

However, many of the kinds of content that Government wishes to address are legal, such

18 Wendy Grossman (2016) *The 0.06 percent* https://www.pelicancrossing.net/netwars/2016/08/the_006_percent.html

19 Rupert Neate (2018) *Over $119bn wiped off Facebook's market cap after growth shock* Guardian https://www.theguardian.com/technology/2018/jul/26/facebook-market-cap-falls-109bn-dollars-after-growth-shock

as clearly spurious "threats", or legal except as part of a pattern or course of conduct, when it becomes harassment.

In cases of unwanted but legal content it is difficult to see what reasonable action the Government can take. In particular, incentivising removal of *legal* content seems a peculiar policy goal. This is exacerbated by the fact that platforms tend to already have restrictions on broad categories of legal content under their community guidelines or terms and conditions.

Platforms typically disallow a range of legal content under their community standards. It is therefore possible for governments to create incentives for platforms to censor additional legal content that is disallowed by platforms, for instance if the government asks platforms to act against 'extremism'.

iv. Technology as a policy instrument

DCMS states that it wishes to "harness the ingenuity of the tech sector, looking to them for answers to specific technological challenges".[20] Technology, such as machine learning, can identify and match patterns and even find approximations that may indicate contextual factors.

However, machines are not yet able to make human judgements about cultural and legal contexts. Machines instead use proxy information to make probabilistic decisions. For some decisions, such as copyright infringement, this will sometimes be highly accurate, for instance in finding literal copies, and otherwise very poor, for instance in deciding if something may be fair dealing, such as a parody. Even literal copies may sometimes mean different things according to context.

Furthermore, technologies are likely to evolve better around broad detection than finessing errors, if the incentives are about detection and removal. Technology can have a role, but policy makers should be clear about its limits.

v. Obligations to monitor and remove

The European Union is currently finalising a new Directive on Copyright in the Single Market.[21] One of the initiatives is to implement Article 17[22] which would impose a filtering obligation on online platforms to scan user-submitted material for likely copyright infringement.

This setup will lead to similar problems as already mentioned in other parts of this document. To ensure compliance and avoid penalty, platforms are asked to err on the side of caution and overblock uploaded content. Content is notified by rights holders, and must then be consistently removed if it reappears ("notice and staydown"). The practical means of achieving this is "upload filtering" or content matching. However this approach will fail to account for exemptions to copyright such as parody, commentary or research. As a result, this will put a lot of strain on freedom of expression.

Additionally, online platforms are likely to tighten their terms of service to be able to delete any content they see fit, even the content is not required to be removed by law. This arrangement will put too much power into the hands of online platforms who will not be required to provide binding ways to appeal their content removal decision.

The EU is considering a similar approach in other areas, such as terrorism, which is problematic for the same kinds of reasons.[23] We should not rely on automated measures alone to identify content that depends on context to judge.

vi. Duty of care

Central to the current proposals is the idea of a 'duty of care'. The idea of a 'duty of care' was mentioned in the consultation questionnaire for the DCMS Green Paper[24] and now in the White Paper. This is elsewhere compared with health and safety and environmental legislation, for instance. However, it is not obvious that a duty of care approach can be simply applied to Internet platforms without significant free expression impacts.

Duties of care are based on the notion of risk management. They are found in health and safety, or environmental legislation. An owner of a physical space or the provider of a service might directly create risks for those using it if they are not sufficiently careful, for instance to maintain buildings or prevent entry to dangerous areas.

These are normally risks which the owner can directly control, and are not about the actions of third parties. This is recognised by proponents, who are clear that they are having to extend the traditional notion of a duty of care considerably

20 https://www.gov.uk/government/publications/digital-charter/digital-charter

21 https://eur-lex.europa.eu/procedure/EN/2016_280

22 https://juliareda.eu/eu-copyright-reform/

23 https://ec.europa.eu/digital-single-market/en/news/public-consultation-measures-further-improve-effectiveness-fight-against-illegal-content-online

24 https://www.gov.uk/government/consultations/internet-safety-strategy-green-paper See

beyond where it has previously been applied.[25] Duties of care have never, to our knowledge, been applied to speech before.

The conflict between a risk approach and the rights of users is easily seen, although this is not discussed in the White Paper at all. For instance, the goals of actors may be directly in competition. Many of the possible examples of 'risk', such as harassment, bullying, drug promotion, or intellectual property infringement, involve multiple parties with potentially different views of their behaviour. Each party is then owed a duty of care. Additionally, the online behaviour may be tangential to some offline behaviour where the real risks play out directly. 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.

There is no easy parallel for the regulation of these harms with particular public or private spaces. Rather, current regulation treats each of these as concerns in which different actors may dispute certain behaviours. There are no obvious examples in which law breaking or bad behaviour becomes the private concern of a private body in order to regulate what is seen as a public risk. The nearest examples might be the conditions placed on clubs and bars to deal with alcohol and drug abuse. However, here again the criteria for harm are relatively easy to distinguish and do not involve adjudicating disputes between parties. A club or bar would not normally have a direct responsibility for any speech or act done by its customers to each other.

The *duty of care* approach may be attractive to government but has a great number of dangers of causing over-reach at platforms. The European Union considered the potential for a 'duty of care' applying to Internet platforms in 2016 in relation to intellectual property rights.[26] It was welcomed by rights holder groups.[27] The proposal was not however advanced as part of the Digital Single Market proposals; it appears to have been too problematic to define. The IT industry considered that it would cause considerable problems within EU law, both in respect of the e-Commerce liability protections and the Charter of Fundamental Rights.[28]

EDRi highlighted the conflicting priorities the *duty of care* approach appeared to create a year later when the Commission again considered pushing the concept:

> **astonishingly, the draft Communication suggests that we need to avoid making undue efforts to make sure that the (possibly automatic) removals demanded by these non-judicial authorities are correct: "A reasonable balance needs to be struck between ensuring a high quality of notices coming from the trusted flaggers and avoiding excessive levels of administrative burden", the leaked Communication says.[29]**

One recent example of a liability aimed at creating a 'duty to act' – reasonably similar to a general duty of care – going badly wrong exists in the recent US provisions to prevent platforms being used for sex trafficking in the SESTA/FOSTA package. The approach that platforms have taken is that *any* activity related to sex working is now disallowed as a potential liability. The result for real-world safety is negative. Many sex workers have relied on online tools in order to increase their personal safety; those that have moved to street working will be at increased risk of rape and attack.

The result for US regulators is also negative. For 125,000 sex workers and clients, they have simply chosen to use an Australian-based equivalent service set up for them specifically, known as Switter.[30] This is less likely to respond to US legal requests and operates on the assumption that it is not subject to US law.

Another 'duty to act' approach exists in Germany in relation to their law compelling platforms to take action when certain laws may be broken.[31] Companies face fines if they do not remove "manifestly" illegal content or illegal content, or face fines of up to €5-50 million. Decisions may be given to a self-regulatory body for an 'independent' review, presumably at cost to the platform. The overall effect is to further

---

25 *Internet Harm Reduction*, William Perrin and Lorna Woods January 2019 https://www.regulation.org.uk/library/2019-Carnegie-Internet-Harm-Reduction.pdf

26 See http://www.politico.eu/wp-content/uploads/2017/09/n-and-a-communication.pdf and https://edri.org/leaked-document-does-the-eu-commission-actually-aim-to-tackle-illegal-content-online/

27 http://ec.europa.eu/information_society/newsroom/image/document/2016-7/fesi_comments_tackling_illegal_content_online_and_the_liability_of_online_intermediaries_13982.pdf

28 https://ecommerce.blogactiv.eu/2016/06/03/does-europe-need-a-new-duty-of-care-for-online-platforms/

29 https://edri.org/leaked-document-does-the-eu-commission-actually-aim-to-tackle-illegal-content-online/

30 https://switter.at

31 Act to Improve Enforcement of The Law on Social Networks, see translation at https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.pdf?__blob=publicationFile&v=2 and Article 19's commentary https://www.article19.org/resources/germany-act-to-improve-enforcement-of-the-law-on-social-networks-undermines-free-expression/

privatise legal decision making, and to incentivise the removal of content.[32]

While a process of negotiation to establish agreed risk mitigations may seem softer-edged than these US and German approaches, legal liabilities for failures to prevent harm (that is, delete content) are clearly core to the proposal. It is hard to see the dynamic towards over-policing not being replicated.

In any risk-based approach, the key question will be the manner in which risk is established, the kinds of risk addressed, to whom, the potential mitigations identified and the proportionality and speech impacts evaluated. In the current proposal, there is no detailed discussion of these aspects, beyond asserting that they will need to be dealt with by the Regulator.

This is particularly acute because 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. However, if the standard for establishing risk is easy to reach, in order to make it easier to compel action, then the risk of disproportionate action and over-censorship increase.

A statutory approach to risk may be very hard to develop in a way that satisfies proponents and avoids over-reach. Under-action or over-action are likely to attract criticism, which ultimately would be the responsibility of legislators to resolve, since they have initiated the framework. Since problems would be unlikely to disappear, ministers would have to explain why their framework had resulted in the problems that emerge from bad actors using the Internet continuing to result in harm. This could produce further political pressure for unrealistic goals of harm elimination.

In contrast, policies developed by platforms within an independent framework may be easier to justify, modify and balance; and errors would continue to be the responsibility of the private actors.

The White Paper singles out risks to vulnerable groups. Elsewhere, there has been discussion of 'harms to society' rather than harms to individuals. Evaluations on either basis contain risks of over-action. As with evidence, the wider the set of risks that are brought in, the higher the scope for the process failing to be legitimate.

We are skeptical that mitigations can be created to make a general and wide duty of care approach fit with the human rights considerations that must be the basis of any changes. It does not seem to us "straightforward in principle" for the simple reason that any *duty of care* is owed to all parties in a dispute. Furthermore, trying to apply a *duty of care* approach to all parties presents further problems, as it would be much harder to assess that duty in relation to persons deemed potentially harmful than to assess the question of harmful content.

vii. Approaches to child safety

Child safety is an important concern for any government. Considerations for children's online safety are particularly fraught, because of the range of issues that children and young adults face at differing ages. It is important that policies are practical and focus on empowerment of children, if they are to succeed. Education, discussion, good parenting and help for under 18s to manage their own risks are more important than any other kind of intervention, as online risks cannot be eliminated. Yet these are not the focus of government policy, which at least rhetorically seems aimed at removing online risks. As a result, many of the policies being pursued in this area seem likely to have little impact on child safety, including age verification.

It is also important to recognise that many potential policies for child safety are in practice restrictions or cause risks to adults, as age verification may do. Filters for instance, when applied to large groups of adults by default, create restrictions which are unnecessary and disproportionate.

To the extent that technical solutions are designed to help children, these can be targeted either towards the child, or at regulating the Internet in general. Technologies that help the child directly are more likely to be useful and effective and avoid the obvious pitfalls of more generalised approaches.

There are particular dangers of disproportionate policy responses if Internet regulation is aimed at making all Internet sites and content safe for children. At present, most Internet sites are not designed for children, but assume their audience is adult. Government should be cautious about assuming that it is possible or desirable to regulate away the risk of children encountering inappropriate material.

32 https://www.article19.org/wp-content/uploads/2017/09/170901-Legal-Analysis-German-NetzDG-Act.pdf

viii. An Internet regulator and co-regulation

The purpose of an Internet regulator appears to be to find policies and approaches to removing certain behaviour or content, on the basis of a notion of harm or of a duty of care. The idea has been put forward in the current White Paper. However,it has not explained how harm or a duty of care might be established in practice, except to assert that it can be, and can be balanced with free expression concerns.

An obvious objection to this approach is that it places any such balance of alleged harms and free expression or privacy risks in the hands of a regulator, to construct a policy with its stakeholders, predominantly the platforms. Whatever the underlying principle that is meant to be applied, a regulator would acquire powers to define the practical limits of speech even where that speech is lawful. This is especially true given the proposed pervasive scope of the White Paper proposal.

Approaches that place significant limits on speech are always problematic. The current government is prevaricating about a 'state regulator' in relation to the press because many of its supporters are reluctant to allow the state to intervene in news publishing. Ofcom's role in placing limits and duties in relation to broadcasters is said to be justified because of the concentrated power of broadcasters to shape public opinion. Now, a regulator is being asked to *ensure* that Internet platforms develop concentrated powers to shape what the public receive.

Given the concerns expressed about the potential power of a state press regulator that would place limits on the free expression of newspapers, better explanation needs to be put forward to explain why a state regulator to limit the free expression of individual citizens is less concerning.

Recommendation 5

Focus online harms policy on process and accuracy rather than hard to quantify and identify risks

IMPROVING REGULATION AND RESOLVING COMPLAINTS

a. Government responsibility for the law

Governments enforce laws: companies comply with their duties. Government can legitimately pursue better enforcement of laws online, but must also accept responsibility for creating processes that allow users, both complainants and posters, to have access to independent processes and an effective remedy for wrongful removal of content.[33] Government must also remember that any response from industry will be a compliance response, and not in any way a step towards enforcement of laws, which is the job of investigators, regulators and the courts.

There are a variety of different scenarios where there are problems for people complaining or having material removed.

New processes need to take into account:

i. **The harm to an individual arising from content remaining available**
ii. **Legitimate aims such as national security**
iii. **The need for an individual to be able to challenge a decision**
iv. **The ability and independence of the person to judge what action to take**
v. **Whether the dispute is a civil matter or a criminal matter**
vi. **Whether the dispute relates to UK laws or breaches in terms and conditions**
vii. **The incentives of each of the actors**
viii. **Any scope for abuse**

33 See for instance *Regulating speech by contract*, Article 19 (2018) https://www.article19.org/wp-content/uploads/2018/06/Regulating-speech-by-contract-WEB-v2.pdf

4

Courts and judicial processes offer the best guarantee of independence. However, any system for challenging content restrictions should also be accessible and inexpensive. Not all decisions need to progress to courts, especially if other fair processes are available. In this chapter, we suggest options that are more likely to satisfy free expression considerations and the rights of all users.

b. Independent self-regulation or co-regulation

It may be appropriate for the larger companies to lead an independent self-regulation effort to be clear about the standards and processes they are putting in place. This could include elements such as alternative dispute resolution for certain types of complaints. Transparency about processes and independent procedures based on human rights standards could deliver improvements for everyone.

Independent self-regulation has the advantage of being independent of government. There may still be negative impacts on free expression however. While privacy rights are to an extent protected by law, restrictions on free expression flowing from private agreements are considered a private matter. Thus self-regulation may be difficult if it attempts to make restrictions on content largely consistent across platforms.

This approach recognises that laws are best *enforced* by governments. Legal *compliance*, rather that enforcement of laws, is the role that private companies can be expected to perform. Self-regulatory efforts are not the same as enforcement of laws.

The government should consider co-regulation as an option for Internet regulation. This would specify the standards for a regulator, such as independence, having regard for free expression, and so on, but the regulator itself would be independent of both government and the Internet companies.

These approaches have greater potential to work internationally, if other countries opt for a similar model. It is more likely to be able to regulate for a wider range of problems and preferences. It may be more likely to achieve public confidence than a state institution, as it is less likely to be seen as government regulation aimed at the backdoor censorship of legal content.

Recommendation 6

Adopt Co-regulation or Independent Self Regulation as the policy model for Internet social media platforms

c. Notice and counter-notice

Notice and counter-notice systems offer the possibility of removing content at scale. Because notice under the E-Commerce Directive normally creates liability, notice and takedown procedures need to be created by law, as with defamation in the UK.

Despite some complaints from copyright holders, notice and takedown has offered an effective mechanism allowing complaints to be handled at scale. Furthermore, it has the potential to be a fair process, as it offers end users the possibility of complaining.

In fact, a major criticism of notice and counter-notice has been that the notices themselves are over-effective. They may be poorly formed, failing to properly meet the requirements of the DMCA (for instance explaining who owns the copyright, and their legal contact details for any counter-notice) or failing to correctly identify the copyright material in question. Additionally, many notices seek to remove material that may successfully rely on copyright fair use in the USA.

Some studies have indicated that on content platforms only 1% of notices are contested, while up to 36% of the notices may be questionable, by failing to properly identify the content, or notifying uses which may be legitimate.[34]

The problem appears to be that users are disinclined to counter-notify as this has the prospect of the other side initiating legal action. Even where they may be confident that they are in the right, this is a daunting prospect. Often the content has little financial value to either the user or the copyright owner, the impact of removal being personal and emotional.

A recent study modelled a notice and counter notice system which allowed users to invoke a dispute resolution with a small cost, where a bad decision would cause a 'cost' to the platform, resulting in more complaints and more accurate decisions.[35] This shows that expert dispute resolution systems could be a useful mechanism to improve notice and counter-notice, especially if combined with incentives to ensure that poor notices are not made. Crucially, end users must not be dissuaded from making a complaint.

Recommendation 7

Application of Notice and Counter-notice systems to content removal procedures

d. An ombudsman

A less heavy-handed approach than an Internet regulator attempting to address all content policies across platforms could be to create an ombudsman that could adjudicate or investigate when specific problems had arisen. This approach normally applies to failure to deliver a service correctly, for instance in relation to ISP service provision.[36] These problems would need to be well-defined, but could help in specific cases where the platform appears to have breached its contractual arrangements with a user.

e. Alternative dispute resolution

Alternative dispute resolution can be considered for some disputes. It is a means for parties to settle disputes by agreement with the assistance of an independent third party that is a lesser authority than a court. It can include arbitration, conciliation, mediation or negotiation as appropriate.

In some cases, ADR could be a very helpful step, especially to resolve incorrect copyright or defamation complaints, if it is risk-free for those who have received notices.

However, ADR is not a magic bullet, and may not solve all issues and complaints. Because terms and conditions are the underlying agreement for most Internet users and content, ADR could create a mechanism for legal content to be removed. This should not be an object of government policy, so should be carefully avoided for instance by judging the limits of expression against human rights standards and appropriate laws.

Any proposal must also ensure that there is an adequate legal process, which could include options for dispute resolution. The State has a duty to balance the competing interests, and competing rights, of different actors and to guarantee the right to freedom of expression. Alternative dispute resolution procedures are not a replacement for court hearings, if either party disagrees and wishes to pursue the matter.

Our comments in this section necessarily outline very top level concerns and remarks.

i. Disputes about legal content that may breach terms and conditions

These are disputes between a platform and their user

Many disputes are about breaches of platform's terms and conditions, where an individual has published something legal that the platform finds unacceptable,

34 Summary in Fiala and Husovec (2018), p4-5

35 Fiala and Husovec (2018) *Using Experimental Evidence to Design Optimal Notice and Takedown Process* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3218286

36 https://www.cedr.com/consumer/cisas/ Communications and Internet Adjudication Scheme

or where an individual believes someone else's material should be removed under those terms and conditions.

Users will always be at a disadvantage when content is removed under terms and conditions. They may not be familiar with the precise delineations, they do not make the initial decisions, and even if a process is fair, they must be determined in order to pursue any complaint. Content is often most relevant at the time it is posted, so gaining the right to put material back sometime later may be a rather limited victory.

There is an imparity of bargaining power. Platforms set the terms and interpret them. Disputes about terms and conditions are resolved solely by the platform, which in practice sets the contract and can create very broad, arguably unfair conditions. Courts in the US and EU so far have been unwilling to engage on the fairness of platform's contracts.[37] Yet in principle, the larger and broader the audience, the more permissive platforms must be in terms of the subject matter they accept, because they are de facto public spaces where limits to speech rights become extremely meaningful. For this reason the UN Special Rapporteur on Free Expression has emphasised the need for platforms to adhere to human rights standards.[38]

To resolve these disputes fairly, users of major platforms need to know that minimum standards are present, for instance clear and predictable rules as to the broad categories of content that is allowed or disallowed, procedures and timetables for review, and a recognition of the relationship of a platform on free expression related to its size and usage. The larger the platform, the more varied the likely uses and the greater the impact that any restrictions create. Thus it is important for the largest platforms like Facebook and Twitter to be more liberal in the content they allow.

Additionally, the platform is the only party in practice that is able to interpret meaning of the contract, often doing so with secret moderation criteria, such as the leaked Facebook moderation handbook.

For instance, if Facebook judges that a link featuring nipples represents nudity that breaches its terms and conditions, it may remove it. Although Facebook now allow appeals, this is a further internal assessment.

Some forms of bullying and 'bad behaviour' may fall into this category, where two users are in dispute about an activity that is lawful, but potentially in breach of terms and conditions. This is a problematic area, as enforcement of behavioural norms can easily infringe people's free expression rights. In whatever way the rules are designed to accommodate these dilemmas, though, the question is one of interpretation of the contract, if the content is otherwise legal. The ability to resolve the meaning of the contract in these disputes currently lies solely with the platform.

In these cases, users are severely disempowered. Platforms should outline what rules on content they have, how decisions are reached and not be able to award themselves the power to remove content arbitrarily. There should be a review procedure when content is removed, and a means to get an independent decision. As a first step, independent and transparent review procedures could be put in place by platforms. Such procedures should be low or no risk for the complainant.

However, if the government requires a specific intervention or procedure such as ADR that could impact free expression, this procedure must be based on restrictions in UK law and human rights standards, rather than the terms and conditions of a platform. Otherwise, the process would be create a means for governments to pressurise companies into restricting speech within terms and conditions, and then to enforce that restriction through a legal process. The result would be a backdoor and extensible means to curtail the use of platforms for legal purposes.

For this reason any independent decision would need to be decided in light of international standards on freedom of expression and domestic law, rather than terms and conditions of a platform.[39]

ii. Private disputes about potentially illegal activity such as harassment

Many disputes between platform users may not be easily resolved by platforms where, for instance, full context of a dispute is not available. This would often be the case in criminal harassment cases, for instance. The dispute may focus on one or more platforms, since harassment is offence based on a course of conduct which may take place in several online venues.

In these cases, both victims and people (potentially wrongly) accused need the ability to get a fair decision

37 https://www.article19.org/wp-content/uploads/2018/06/Regulating-speech-by-contract-WEB-v2.pdf *Regulating Speech by Contract*, Article 19 See pages 15 and 37

38 https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/Privatesectorinthedigitalage.aspx *Freedom of expression and the private sector in the digital age* Kaye 2016; *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*, 6 April 2018 https://freedex.org/a-human-rights-approach-to-platform-content-regulation/ See paras 64-72

39 See Article 19 (2018) *Side-stepping rights: Regulating speech by contract* p 38 Recommendation 3: Right to an effective remedy between private parties should be provided

easily, without this being the sole determination of the platform. An arbitration procedure could be created for the most civil common private disputes. This cannot include criminal matters. Those must be pursued through the courts.

A further limitation is that parties in private, civil disputes will often not wish to use dispute resolution mechanisms, but prefer to go to court. If dispute mechanisms are a precursor to court action, there is some danger that they will become a formality.

iii. Private disputes about copyright and defamation

As noted above, notification of copyright infringement at a platform results in removal of content in UK law. In contrast, in US law, users can issue a counter notice, although in practice most people do not, as it may result in court action. As a result, legal material is removed.

In defamation, the UK has a similar system that allows reputation management companies to ask for material to be removed, unless the poster agrees to hand their personal details to the company. This will often be successful, as posters do not want to engage in a legal dispute, even when they know the claim is spurious.

In these cases, a zero or very low risk system of arbitration would allow users to contest spurious complaints. Such a system could also be designed to identify bad actors, who for instance consistently fail to ensure they are issuing accurate complaints. Incentives could be introduced to reduce abusive behaviour.

A further step could be to allow individuals to take legal action against persons making spurious copyright complaints, such as exists for baseless patent threats.

iv. State actors seeking to resolve potentially illegal content or activity

These are disputes between a government agency and a private individual, such as a user on a platform or domain name owner

Government and public bodies require a more robust process to restrict content, as their powers are more sweeping and have greater effects. They also need to be accountable for the way they use their powers. Where no prior court decision is required, a system of notice and counter-notice should always take place in order that users can assert and defend their right to publish, unless there are very clear and exceptional reasons otherwise.

We propose an independent and impartial judicial authority dedicated to handling take takedown requests at Nominet, CTIRU and IWF. This should review requests prior to action at Nominet and by CTIRU. Nominet and CTIRU should adopt a notice and counter notice system to allow individuals to dispute any claim they make. The independent authority should then review individual complaints made.

Recommendation 7

Create independent arbitration processes to review complaints at platforms based on terms and conditions and community standards

Recommendation 8

Create arbitration processes to resolve private disputes about unlawful activities at platforms

f. Summary table

NATURE OF DISPUTE OR ISSUE	APPROPRIATE APPROACH	LIMITATIONS
Legal content in breach of terms and conditions, including inappropriate but legal behaviour	Transparent procedures and independent appeals within each platform. Improved accessibility to procedures. Independent self-regulation or Co-regulation to enhance accessibility, responsiveness and accuracy. Ombudsman could help with accessibility and quality. Decisions are still at companies.	The state cannot legitimately instruct others how to deal with legal content or behaviour. An ombudsman should not be used to enforce contractual limitations on speech.
Private disputes about potential civil or criminal wrongs, such as defamation or copyright infringement.	Notice and counter-notice. (takedown except where the notice is disputed). ADR to resolve disputes, but must be very low cost to access.	Notice and takedown systems are open to abuse unless appeals are low risk or no risk.
Private disputes about potentially unlawful behaviour such as harassment.	Accessible legal procedures to hold individuals to account. This could include use of behaviour orders.	Not all of these disputes will reach a threshold of criminality or anti-social behaviour. Behaviour may occur on multiple platforms or with multiple accounts. Individual platforms cannot restrict individuals' activities, but instead act against accounts or content specifically.
Stage or other agency content removal requests of potentially illegal material	Prior independent review of requests. Independent review of complaints.	Agencies desire easy and quick procedures.

RECOMMENDATIONS TO GOVERNMENT 5

Recommendation 1

Increase the capacity of Government to deal with Internet issues

Recommendation 2

Increase the capacity of Lords and Commons to conduct detailed research and policy work

Recommendation 3

Incorporate the No General Monitoring obligation into UK law

Recommendation 4

Application of Notice and Counter-notice systems to content removal procedures

Recommendation 5

Focus online harms policy on process and accuracy rather than hard to quantify and identify risks

Recommendation 6

Adopt Co-regulation or Independent Self Regulation as the policy model for Internet social media platforms

Recommendation 7

Create independent arbitration processes to review complaints at platforms based on terms and conditions and community standards

Recommendation 8

Create arbitration processes to resolve private disputes about unlawful activities at platforms

HARMS SUMMARY

Specific area of harm and concern

ISSUE	ISSUE DESCRIPTION: MAIN FEATURES OF THE PROBLEM QUESTIONS THAT ARISE
ILLEGAL CONTENT OR ACTIVITY	
Terrorism content	Some terrorist content is easy to spot because of branding or multiple reposting. Major platforms have taken action to reduce and remove material via automated detection, e.g. Tech against Terrorism initiative: https://www.techagainstterrorism.org/ Other content is potentially difficult to identify especially when conversations take place through private and encrypted communication channels. There are circumstances where accessing and viewing such content is entirely justified, such as research and journalism. Some content e.g. footage of conflict zones, can also get caught by filters but be legitimate to post, and removing this content too quickly can frustrate legitimate access.
Radicalisation / Extremism	Radicalisation is a process whereby an individual comes to embrace values and opinions about a certain topic that gradually become more extreme while at the same time finding it more difficult to accept opposite opinions. Facilitated by online access due to uncensored messages, echo chambers and a sense of anonymity in what is viewed. Young people seeking community and acceptance online are particularly susceptible to messaging. Addressed in part by platforms through the UK counter-terrorism and CVE programmes Prevent and Channel. Potentially difficult to intercept, especially when conversations take place through private and encrypted communication channels.

ISSUE	ISSUE DESCRIPTION: MAIN FEATURES OF THE PROBLEM QUESTIONS THAT ARISE
Child abuse images	General consensus exists that images should be removed. Major platforms are proactive in doing so - see e.g. Internet Watch Foundation: https://www.iwf.org.uk Can be difficult to monitor where image sharing takes place through encrypted channels or direct file-sharing platforms. Third-party apps for discovering WhatsApp groups allow for the trading of images of child exploitation. Even so, appeals mechanisms are necessary as mistakes are made. Implementation, for instance of blocks or domain suspensions, can lead to problems.
Grooming and Child Sexual Exploitation	Grooming is a course of conduct where an adult attempts to gain the trust of a minor to facilitate abuse. Potentially difficult to identify especially when it takes place through private communication channels, where messages are encrypted. Estimating prevalence is problematic due to the unreliability of official estimates and given that self-report surveys are reliant on the willingness of young people to disclose abuse. Platforms are taking some action to empower young users by recommending steps they can take to protect themselves online, but there is no “privacy by design”. Understanding whom children are likely to confide in when distressed about online sexual solicitation is vital as there has been limited research conducted in this area.
Illegal speech • **threats** • **hate speech** • **statements of criminal intent**	Some hate speech, where it promotes violence or active prejudice/ discrimination against groups of people, reaches a criminal threshold. However, platform takedowns can be arbitrary or ineffective to counter real-world harm. Where material is removed, appeal mechanisms are necessary as complex situations can lead to disputes.
Revenge pornography	Non-consensual posting of sexually explicit images or video by a former intimate partner. Increasingly subject to criminal liability globally. Continued presence of material online causes ongoing emotional harm. Material itself can remain legal at all times and activity of posting is only unlawful after conviction.
Knife sales	Can be difficult to judge images skirting the bounds of legality.
Drugs sales	Content relating to illegal sales is generally removed or reported to police. Sales could look legal if coded messaging is used.

ISSUE	ISSUE DESCRIPTION: MAIN FEATURES OF THE PROBLEM QUESTIONS THAT ARISE
Sale of legal but potentially harmful items	Includes acid and other chemicals. Also includes innocuous items such as sage where these promise a cosmic outcome. Context is key but complex.
Hacking and Cybercrime	As young gamers join hacking forums to source game modifications, they can be 'groomed' by cybercriminals who recognise their skills and attempt to exploit them by encouraging them to participate in illegal online activities. Games may be riskier for vulnerable children such as those with special educational needs insofar as they find it difficult to judge what is real or to read the intentions behind an approach by other players. Children (hackers are mostly boys) often engage in hacking activities for fun without realising the criminal consequences of their actions. Online deviance such as digital piracy is often minimised, since the internet is perceived as a place with no guardians or laws.
LEGAL BUT POTENTIALLY HARMFUL CONTENT OR ACTIVITY	
Hate speech • **Anti semitism, racism** • **Misogyny** • **Homophobia**	Offensive speech not reaching bounds of illegal hate speech can cause emotional harm and a toxic online environment for certain categories of user. Content may be offensive but protected by free speech rights.
Harms to democracy • **disinformation and 'fake news'** • **(attempted) interference with election processes** • **intimidation of politicians and those engaged in politics (e.g. candidates or other public figures)**	Disinformation and 'fake news' is generally legal, intimidation and interference are generally illegal. Interference content can be entirely innocuous on its face but be potentially harmful because of the source and overall intent (e.g. a foreign power nefariously subverting democratic processes by embedding a societal perception of 'us' vs 'them'). Direct intimidation of political figures falls within the hate speech categories above. Impact is affected by people's position in society and the wider anti-democratic silencing effect of intimidation.
Harms to justice: interference with criminal proceedings and the trial process	This can include sharing confidential information about ongoing trial proceedings, inappropriately contacting or attempting to contact actors in the trial process, including witness intimidation.

ISSUE	ISSUE DESCRIPTION: MAIN FEATURES OF THE PROBLEM QUESTIONS THAT ARISE
Harassment and trolling	Whilst aggregation of content online in a way that is not feasibly possible in the physical world can cause emotional harm, removing content can violate free speech if discrete postings are all legal and originate independently from multiple users and harm caused does not reach a recognised criminal threshold.
Self-harm and suicide content	Self-harm is a complex topic. Technical controls exist for blocking such content through home network-level filters but real-world support systems are also needed. Removing self-harm images can lead to social isolation and increased offline harms - talking about self harm is not illegal and can be helpful for people processing personal issues. Online communities dedicated to self-harm and suicide act as support systems for excluded and marginalised children by providing them with peer support and positive identity formation. Self-posted content online can also lead to people receiving help if others see that they are self-harming, for instance.
Promotion of anorexia/ unhealthy body image	There is a fine line between promotion and self-expression, particularly within community groups. People posting about body image may be seeking help, which content removal would stifle.
Drugs promotion	Drugs promotion can range from lifestyle sites, sales of legal drugs 'paraphernalia', to discussions of legalisation that come from groups who already use drugs.
ISSUES SPECIFIC OR PRIMARILY AFFECTING YOUNGER AGE GROUPS	
Cyberbullying and trolling	Bullying comprises a wide range of behaviour, from micro-aggressively "liking" posts, making consistently nasty comments, trolling, posting pictures without consent, posting information that can identify an individual's location (potentially then leading to physical offline attacks), to targeted abuse and threats. A large amount of cyberbullying is based on identity-related characteristics (i.e. appearance, sexual activity, religion, gender, race/ethnicity, disability). Children receiving unwanted sexual attention from adults is also a form of cyberbullying. Prevalence and impact varies by age, gender and sexual orientation.

ISSUE	ISSUE DESCRIPTION: MAIN FEATURES OF THE PROBLEM QUESTIONS THAT ARISE
Sexting and sexual harassment	This can be conceptualised as part of cyberbullying and online harassment more generally but it has specific gendered aspects that make it a distinct form of online victimisation (e.g., slut-shaming, homophobic comments). The wider context matters – the prevalence of gender inequalities, sexual stereotypes and coercion, and a lack of understanding of consent all serve to blur the boundaries between sexting and harassment; ultimately, girls are more at risk. Experiences are often associated with developing intimate relationships as teenagers. Where harassment involves peers, this can lead to offline harassment, generally in school settings, leaving young people feeling trapped and unable to escape these experiences.
Exposure to sexual and violent images	Pornography and violence seen most often on video-sharing sites, followed by other websites, then social networking sites and games. Unintentional viewing of pornography can happen via pop-ups, misleadingly named websites or advertising on illegal streaming sites. Exposure to pornography adversely impacts children's sexual attitudes, expectations and beliefs, particularly through developing unhealthy attitudes towards women. Viewing of other sexual and violent images ranges greatly in terms of impact and what is appropriate for any particular child. "Sexual images" could include intimate surgery videos, which have a legitimate medical interest, and may be positively sought out e.g. by LGBT+ teenagers exploring identity.
Other age-inappropriate material	Includes dangerous viral challenges and prank videos. Can also include fakes e.g. MOMO, disturbing videos that are labelled to appear harmless e.g. the Peppa Pig spoof videos, and calls to engage in dangerous activity. Could include swearing. Ranges greatly in terms of impact and what is appropriate for any particular child.
Gangs	Often closely linked to grooming. Can be difficult to draw distinctions between gangs and friendship groups.
Addiction / overuse	Technology system design can facilitate, encourage and amplify the above behaviours/harms. It can also lead to overuse and associated disrupted childhood harms (e.g. addiction, anxiety, aggression, sleep deprivation, memory impairment).

Written by Jim Killock, Executive Director, Open Rights Group

Open Rights is a non-profit company limited by Guarantee, registered in England and Wales no. 05581537

Open Rights Group, Unit 7, Tileyard Acorn Studios, 103-105, Blundell Street, London, N7 9BN

Registered Office: 12 Duke's Road, London WC1H 9AD

www.openrightsgroup.org/contact/

www.openrightsgroup.org