INDEX ON CENSORSHIP

A VOICE FOR THE PERSECUTED

RIGHT TO TYPE

How the "*Duty of Care*" model lacks evidence and will damage free speech

Introduction: The Duty of Care puts a fundamental freedom at risk

Since the abolition of press licensing in 1695, newspapers in England have been free to print without direct government interference. This principle that the written and spoken word should not be subject to specific censorship by government, rather that a series of laws (from defamation to public order) should apply equally to all, has been the underpinning of our right to free speech.

This freedom is in peril.

The academic concept that lies behind the government's draft Online Safety Bill will erode this freedom. With little debate, an abstract concept, the "**Duty of Care**" has become central to civil service thinking about freedom of expression online. The "Duty of Care" applies notions best applied to health and safety law in the workplace to freedom of speech online. It will reverse the famous maxim, "published and be damned", to become, "consider the consequences of all speech, or be damned". It marks a reversal of the burden of proof for free speech that has been a concept in the common law of our country for centuries.

Worse of all, this academic concept is about to be applied to the new frontier of free speech, the internet, with little debate or scrutiny through the draft Online Safety Bill. This Bill will fundamentally shift who controls free speech in the UK, from Parliament, to Silicon Valley companies who will be responsible for moderating and censoring content which is, in law, perfectly legal but deemed 'harmful'.

'Legal but harmful' has been defined in the draft Bill as causing "physical or psychological harm", but how can this be proved? This definition opens up significant problems of subjectivity. The reason, in law, we do not use this definition for public order offenses is that it is hard for citizens to understand how their words (written or spoken) could cause psychological harm in advance, especially on the internet where we do not know our audience in advance.

It will be up to Silicon Valley companies to decide the type of free speech that is 'harmful' and censor it accordingly. Yet, the notion of 'legal but harmful' opens up more questions than it answers. What if my definition of harmful differs from yours? How can Silicon Valley executives adjudicate on issues of taste, or offence, without an understanding of the context of British culture, where comedy and the arts often intend to provoke?

Censoring content that is 'legal but harmful', is not lawful. It breaches fundamental principles in the common law and our human rights framework. The intention seems to be to sidestep Parliament.

Even more problematic is the role of Ofcom as the final adjudicator. For centuries, we've enjoyed a judicial not technical process for adjudication over matters of free speech. If you write something that is illegal, you can be held accountable for it. Instead, the draft Online Safety Bill will give Ofcom the powers to fine technology companies for allowing content to be posted that is perfectly legal, but deemed harmful by the subjective standards of the regulator. Ofcom is not entirely shielded from political interference, it is not a flight of fancy to imagine a scenario in which Ofcom could be subject to a push from government to clamp down on free speech that offends the government of the day. The fines that Ofcom can levy are eye-watering, with the potential fines as high as 10% of turnover. For Facebook or Google that would be $8.5 billion[1] and $18.2 billion[2] respectively. There will be a commercial incentive to over-censor, to remove content once deemed as perfectly acceptable, as to defend free speech online could cause significant financial risk.

Ultimately, there is a simple solution that fixes the draft Online Safety Bill: focusing the Bill on content which is illegal under laws passed by parliament. By picking up the policy solutions proposed by the Centre for Policy Studies 'Safety without Censorship' report, the government can deliver on its intention to ensure that the most harmful content is removed in a timely manner.[3] There is an established legal framework to deal with the majority of the harms considered by the major advocates for the draft Online Safety Bill: from grossly offensive content, to content that promotes hate crimes and terrorism, to online sexual abuse. The application of the laws can be delivered with amendments that limit the scope of the legislation to content that is illegal.

By keeping the scope, we can protect free speech, while delivering on the good intentions of the legislation. But, as it stands, the draft Online Safety Bill is too broad, too sweeping and will harm freedom of speech that we have fought so hard as a society to protect.

---

[1] Facebook, 'Fourth Quarter and Full Year 2020 Results' (January 2021), https://www.prnewswire.com/news-releases/facebook-reports-fourth-quarter-and-full-year-2020-results-301216628.html accessed 17 June 2021.

[2] Alphabet, 'Alphabet Announces Fourth Quarter and Fiscal Year 2020 Results' (February 2021, https://abc.xyz/investor/static/pdf/2020Q4_alphabet_earnings_release.pdf?cache=9e991fd accessed 17 June 2021.

[3] Centre for Policy Studies, 'Safety Without Censorship: A Better Way To Tackle Online Harms' (September 2020), https://www.cps.org.uk/files/reports/original/200926202055-SafetywithoutCensorshipCPSReport.pdf accessed 17 June 2021.

Executive Summary

- The Duty of Care is a flawed concept based on limited evidence, which would create a radically new framework for the proactive censorship of entirely legal content.

- The Duty of Care model's 'precautionary principle' shifts the balance from punishing individuals for breaking the law to state sanctioned private censorship of content for which there has been no judgement whether it is legal or illegal. This shift, as enacted through the Online Safety Bill, would mark the most significant change in the role of the state over free speech since 1695.

- The Duty of Care model embedded in the Online Safety Bill will not just impact on the right to publish, but also on the right of UK citizens to receive ideas and information. Previously, the state has had a limited role in determining the content that citizens can receive, with prosecution for content deemed illegal happening after the content has been seen by citizens, Online Harms will give the state the powers to enforce immediate censorship of content that is deemed harmful.

- Ultimately, there is no strong evidence base to justify the algorithmic censorship of content which under the common law is perfectly legal. We believe that if it is legal to say or print, you should be free to type it.

- There is a significant evidence base that the use of algorithmic censorship will disproportionately impact upon marginalised groups, often the groups of people for whom the ability to publish online is most critical as they have few other platforms.

- There are a wide range of existing laws to deal with harmful content, the issue is that there is currently too little police resourcing to deal with the prosecution of illegal content. Where an evidence base exists that there are gaps in the existing legal framework, Parliament should pass or amend laws to deal with these gaps.

- The Online Safety Bill will shift decision making over freedom of expression online from Parliament to private Silicon Valley companies, who will not risk fines as high as 10% of their turnover to defend free speech for ordinary British citizens. The risk of over-censorship is significant.

- The Online Safety Bill carve outs for journalism and politicians will create two tiers of free speech in the UK. The carve out for journalism is limited in scope and will not prevent the blocking of opinion or investigative journalism by algorithms that deem the content potentially in breach of the Ofcom the content potentially in breach of the Ofcom guidelines. In practice, UK newspapers would need to take heed of Ofcom guidance, or see their content removed from social media platforms.

Why will the draft Online Safety Bill and the 'duty of care' model erode freedom of speech?

1. The Bill creates a two tier system for freedom of speech that gives greater freedom to politicians and journalists.

The two-tier system for content moderation censors the masses, and privileges the elites. The Bill contains specific exemptions for journalistic content and content of 'democratic importance' essentially privileging the speech of a select few over the British public. However, even on these terms, the Bill fails to protect online newspapers adequately and the notion that free speech should be treated differently because of who is exercising the right of free speech, is problematic.

The Bill attempts to give politicians and journalists additional protections for their free speech. Yet, the ability of ordinary people to speak their mind online has brought about many positive changes such as the #MeToo movement. The Bill could lead to companies removing people's own experiences as they could be considered 'harmful'. The voices of ordinary British people are not any less important than those of a select few.

Not only are the carve outs unfair, they open up a whole host of other problems. The relevant Minister will have the ability to change the parameters of the carve out at will without consulting Parliament. For example, decisions taken during the Covid-19 pandemic clearly demonstrate the pitfalls of allowing Ministers to take decisions unilaterally.

The exemptions for journalism fail to protect the dissemination of content by trusted online news sources through social media networks. For instance, while news published on the Telegraph's website would be subject to the journalistic exemptions in the Bill, the article disseminated on Facebook would be subject to the algorithmic removal of 'legal but harmful' content. This could lead to the removal of entirely lawful journalism from social media platforms that fall foul of an algorithm attempting to protect users from 'legal but harmful' content. There is no detailed definition of journalism in law, nor despite repeated attempts, including at the United Nations and Council of Europe, has a satisfactory definition of a journalist been found. A wave of citizen journalist claims could slow down the appeals process for journalistic content meaning that news content from major newspapers could be caught up in the backlog and blocked during crucial news moments such as elections or national emergencies. In practice, newspapers may find opinion, even news pieces that cause 'offence' to certain audiences, will face either algorithmic throttling to reduce their potential viral sharing online, or outright blocking.

By creating two tiers of freedom of expression, the government is opening up disparities between the free speech enjoyed by ordinary citizens and that enjoyed by a Westminster bubble of journalists and politicians.

2. If something is legal to say, you should be free to type it.

The very concept of the Duty of Care with the proactive duty to censor online content, means that there will be a disparity between what it is legal to say uncensored and what it will be possible to type online. Speech which is entirely legal to print on a poster, or to place in a newspaper, or magazine, or leaflet, it will not be possible to publish online on a Category 1 social media platform.

This is because the threshold for free speech will be different on Category 1 social media platforms than it is in law. Companies like Facebook will be required to actively and proactively monitor content to ensure it complies with the Duty of Care and does not cause foreseeable harm, this is a deliberately constructed to bring within scope content which is currently perfectly legal (see next section).

Further to this, in our democracy it is Parliament's role to define the parameters of illegal speech and content, and for the courts to determine whether publication is legal or illegal. Social media platforms and the algorithms they will introduce to censor content will not have any due regard for the nuance of the common law, legal defences, or the broader legal framework which defends and upholds freedom of speech. An algorithm will decide on the basis of the words used, or images used, whether content should be taken down. This leaves no space for nuance, irony, satire, political context, or justifications in law from public interest to truth to fair comment.

As noted in the CPS report, there is strong evidence to suggest that online platforms and algorithms designed to detect hate speech over-censor. Social media platforms have a history of removing public interest content. In 2016 Facebook sparked public outrage by deleting the iconic 'napalm girl' photo, one of the enduring images of the Vietnam War, due to nudity concerns.[4] Facebook and Twitter restricted sharing of the New York Post for reporting on allegations surrounding Hunter Biden in the lead up to the US presidential elections in November 2O2O.[5]

---

[4] The Guardian, 'Facebook backs down from 'napalm girl' censorship and reinstates photo' (September 2016), https://www.theguardian.com/technology/2016/sep/09/facebook-reinstates-napalm-girl-photo accessed 17 June 2021.

[5] Vox, 'Facebook and Twitter took drastic measures to limit the reach of a disputed news story about Hunter Biden' (October 2020), https://www.vox.com/recode/2020/10/14/21516194/hunter-biden-new-york-post-facebook-twitter-removed accessed 17 June 2021.

Removing these posts have significant real world implications. Human Rights Watch has highlighted how social media content, particularly photographs and videos, posted by perpetrators, victims, and witnesses to abuses, has become increasingly central to prosecuting crimes and putting criminals in jail.[6]

The proponents of the draft Online Safety Bill have created a new construct, the Duty of Care, that bypasses both Parliament and the Courts, in order to create a new framework for private censorship. That this has happened with so little debate, or scrutiny, is of deep concern and could have major implications both within the UK and across the globe.

3. The Duty of Care is a flawed concept based on limited evidence

The academic concept that lies behind the government's draft Online Safety Bill, the "Duty of Care", was developed by Professor Lorna Woods and Will Perrin for the Carnegie Trust.[7] The "Duty of Care" is based on the precautionary principle, that a regulator should act on emerging evidence rather than with full scientific certainty. The authors state:

> *The regulator is there to tackle systemic issues in companies and, in this proposal, individuals would not have a right of action to the regulator or the courts under the statutory duty of care.*

The application of a concept that underpins health and safety law in the workplace to freedom of speech online, and the removal of individuals (and collective groups) right to challenge decisions in the courts, would mark a fundamental shift away from the individual towards the power of the state. Rather than punishing individuals for breaking the law (laws that have to balance free speech with other rights), the notion of the 'Duty of Care' gives the state the power to set parameters for the proactive censorship of content by companies.

---

[6] Human Rights Watch, ' "Video Unavailable" Social Media Platforms Remove Evidence of War Crimes' (September 2020), https://www.hrw.org/report/2020/09/10/video-unavailable/social-media-platforms-remove-evidence-war-crimes accessed June 17 2021.

[7] Carnegie UK Trust, 'Internet Harm Reduction: A Proposal' (January 2019), https://www.carnegieuktrust.org.uk/blog/internet-harm-reduction-a-proposal/ accessed 17 June 2021.

In their evidence[8] to the House of Lords Communications Committee Inquiry (2018), Professor Lorna Wood and William Perrin gave the following analogy which underpins their conceptualization of the Duty of Care:

> *Many commentators have sought an analogy for social media services as a guide for the best route to regulation. A common comparison is that social media services are "like a publisher". In our view the main analogy for social networks lies outside the digital realm. When considering harm reduction, social media networks should be seen as a public place – like an office, bar, or theme park. Hundreds of millions of people go to social networks owned by companies to do a vast range of different things. In our view, they should be protected from harm when they do so.*

The analogy with a theme park, or bar, is misleading and problematic. We do not hold publicans directly responsible for the speech of their patrons prior to the speech occurring. It would not be reasonable to suggest that a pub should have a proactive duty to prevent offensive speech happening, and that it should be fined a proportion of its turnover if it failed to put measures in place to stop such speech happening.

Fundamental to the concept of the 'Duty of Care' is that speech and content that could cause harm, should be proactively taken down by internet companies. That reduction of the right to publish ideas that may be offensive, but are legal, is a significant reduction in the space for freedom of expression. Yet, the 'Duty of Care' goes further than just the prevention of publication.

The 'Duty of Care' also reframes the relationship between the state and the individual in relation to the receiving of ideas. In modern times, the state has had no role in preventing citizens from reading literature or concepts that could cause 'harm' or 'offense' - from Nabokov's Lolita, to Qubt's Milestones, Rushdie's Satanic Verses or Nietzsche's Thus Spoke Zarathustra. Yet the 'Duty of Care' places on internet intermediaries liability for the dissemination of ideas, if those ideas could be considered to cause harm. In the worldview of Woods and Perrin, the state must sanction the social network if it does not proactively reduce the space for ideas, concepts or content that could cause harm.

[8] Carnegie UK Trust, 'House of Lords Communications Committee Inquiry. Response from Professor of Internet Law Lorna Woods, University of Essex and William Perrin – written evidence' (May 2018), https://d1ssu070pg2v9i.cloudfront.net/pex/carnegie_uk_trust/2019/01/25143759/House-of-Lords-Evidence-Final.pdf accessed 17 June 2021.

While well intended, there is no space in the academic notion of the 'Duty of Care' for the notion that adults may wish to receive or research ideas that could challenge, disturb and cause offence or harm to one's psychological well-being.

The draft Online Safety Bill defines harm as "physical or psychological harm". The harm that the 'Duty of Care' is meant to capture is harm that is currently legal (as decided by Parliament) but causes "physical or psychological harm". As we note below, illegal harm is already subject to a plethora of laws with a low threshold for successful prosecution, as previously pointed out by Index on Censorship in our free speech scorecard.

As Index on Censorship noted in our submission to the Online Harms White Paper consultation,[9] the wide range of harms that the government is seeking to tackle through this legislation requires different responses. Index believes that any measures proposed should be underpinned by evidence of the scale of the harm and the effectiveness of any measures taken. The evidence base for the Internet Safety Strategy Green Paper has been variable in its quality. As we explained:

> *For example, the recent study by Ofcom and the Information Commissioner's Office Online Nation found that 61% of adults had a potentially harmful experience online in the last 12 months. However, this included "mildly annoying" experiences. Not all harms need a legislative response.*

With a limited evidence base, and no certainty around the scale of the problems the Bill is intended to tackle, it is hard to see any justification for the restriction of "legal but harmful" content, rather than new legislation that tackles any gaps in the law that may exist in tackling harmful content online .

'Legal but harmful' has significant problems of subjectivity especially around the definition that such content causes "physical or psychological harm". How can this be proved, and what is the evidence base that Ofcom will work from? How can internet users understand whether their words will cause psychological harm in advance, especially on social networking sites, where content can travel worldwide in almost an instant. How does cultural sensitivity work in this context, should the major social networking sites develop algorithms fit for cultural sensitivities in both Saudi Arabia and the UK?

---

[9] Index on Censorship,'Index on Censorship submission to Online Harms White Paper consultation' (June 2019), https://www.indexoncensorship.org/wp-content/uploads/2019/07/Online-Harms-Consultation-Response-Index-on-Censorship.pdf accessed 17 June 2021.

As lawyer Graham Smith, the author of the canonical Internet Law and Regulation, notes in his longread article on the Duty of Care.[10] the UK Supreme Court judgement in Rhodes v OPO illustrates the legal difficulties in linking the prospect of psychological harm with speech:

> *"It is difficult to envisage any circumstances in which speech which is not deceptive, threatening or possibly abusive, could give rise to liability in tort for wilful infringement of another's right to personal safety. The right to report the truth is justification in itself. That is not to say that the right of disclosure is absolute … . But there is no general law prohibiting the publication of facts which will cause distress to another, even if that is the person's intention."*

As Smith notes:

> *This passage aptly illustrates the caution that has to be exercised in applying physical world concepts of harm, injury and safety to communication and speech, even before considering the further step of imposing a duty of care on a platform to take steps to reduce the risk of their occurrence as between third parties, or the yet further step of appointing a regulator to superintend the platform's systems for doing so.*

There is also no place in the policy framework for the benefits of freedom of expression as a social good. In their evidence to the House of Lords Communications Committee Inquiry, Woods and Perrin see ample opportunity to go even further than a proactive duty of censorship, and also proposed internet companies should give their users tools to block speech they disagreed with:

> *7. a) measures to empower users, for example pre-emptive blocking tools in the hands of the user; setting up sub- groups that have different toleration of certain types of language*

With these policy remedies, internet users could further find themselves in the 'filter bubble' identified by technologist Eli Pariser, no longer subject to viewpoints or ideas that challenge their thinking. The 'filter bubble' has been highly problematic in sensitising internet users to ideas that challenge them, laying the foundations for hyperpolarisation in politics and creating the space for disinformation campaigns by hostile states to take root.

---

[10] Cyberleagle, 'Take care with that social media duty of care' (October 2018), https://www.cyberleagle.com/2018/10/take-care-with-that-social-media-duty.html accessed 17 June 2021.

The Bill's introduction of the "Duty of Care model" which comes from Health and Safety legislation lacks a strong evidence base and would be harmful to free speech. The draft Bill, as intended, would effectively outsource internet policing to Silicon Valley. The new rules would force tech platforms to delete posts that are considered "harmful" but does not define what is and is not "harmful", this will result in many perfectly legal posts being proactively censored. Any platforms found by Ofcom to have failed in their Duty of Care could be subject to fines of up to £18 million or ten per cent of their annual global turnover. In addition, senior management officials at tech companies who fail to meet their commitments could be subject to criminal prosecution - certainly a good motivation for them to apply sweeping prior censorship of content. Whereas the strong application of health and safety legislation can make us safer in the workplace, when it comes to speech it makes us all less free.

4. AI is a blunt instrument on speech that further silences marginalised communities.

The UK government passing the buck to tech platforms to moderate speech will lead to the censorship of marginalised communities. Tech companies will be forced to use Artificial Intelligence to censor posts which are unable to grasp the complexity of human language - especially when it comes to different languages and cultural nuances.

Numerous studies have shown how the existing censorship algorithms on social media platforms disproportionately censor members of marginalised groups.[11] We saw this in 2018, when Tumblr introduced AI for identifying and removing adult content and reportedly routinely misclassifies innocuous material, with content by LGBTQ+ users seemingly particularly penalised.[12] On Instagram, plus-sized and body-positive profiles were often wrongly flagged or shadow banned for "sexual solicitation" or "excessive nudity".[13] While on TikTok internal documents revealed that its moderators had been instructed to suppress posts created by users deemed 'too ugly, poor, or disabled'.[14]

---

[11] Cobbe, J, 'Algorithmic Censorship by Social Platforms: Power and Resistance' (October 2020), https://link.springer.com/article/10.1007/s13347-020-00429-0#ref-CR13 accessed 18 June 2021.

[12] ArsTechnica, 'Tumblr's porn ban is going about as badly as expected' (May 2018), https://arstechnica.com/gaming/2018/12/tumblrs-porn-ban-is-going-about-as-badly-as-expected/ accessed 18 June 2021.

[13] The Guardian, 'Instagram's murky 'shadow bans' just serve to censor marginalised communities' (November 2019), https://www.theguardian.com/commentisfree/2019/nov/08/instagram-shadow-bans-marginalisedcommunities-queer-plus-sized-bodies-sexually-suggestive accessed 17 June 2021.

[14] The Intercept, 'Invisible Censorship: TikTok told moderators to suppress posts by "ugly" people and the poor to attract new users' (March 2020), https://theintercept.com/2020/03/16/tiktokapp-moderators-users-discrimination/ accessed 17 June 2021.

Ethnic minority groups would be disproportionately impacted by these AI algorithms. Academics recently developed 'HateCheck' which tested the efficiency of hate speech detection models using 29 key functionalities.[15] They found that all four of the four models tested with HateCheck had fundamental weaknesses including wrongly blocking rebuttals to hate speech and bias when it comes to certain groups.. An analysis of popular hate speech datasets and classifiers found that tweets by African-American users were up to two times more likely to be labelled as offensive than tweets by others. Researchers also demonstrated that Perspective disproportionately identifies posts written in African-American Vernacular English as "rude" or "toxic," reflecting and amplifying racial bias. Similarly materials in Urdu and Arabic are twice as likely to be removed than English language content.

The algorithms can have a significant impact in ongoing conflicts and crises. In recent weeks both Facebook and Twitter have incorrectly taken down and removed millions of mostly pro-Palestinian posts and accounts, blaming the takedown on their AI software.

5. Criminals should be prosecuted, not simply censored.

Censoring speech online will not tackle the root causes and consequences of harmful behaviour. Tackling illegal content requires greater resourcing to enforce existing criminal laws while education programmes are a better approach to prevent harmful content. Criminals should be put behind bars, it is not enough to delete their content by algorithm.

The majority of abhorrent content online, such as images of child sexual abuse or racist incitement to violence, is already illegal under current laws. However, the police service is under equipped to deal with the volume of online crime. In 2019 the five former top police officers in Britain signed an open letter warning of the dire consequences of the dramatic under-resourcing of the police in recent years.[16] Prior to changes in 2019, only 31% of forces in England and Wales had a dedicated cyber capability and many are still playing catch up.[17] Annual reports by the government have repeatedly found a large cyber security skills shortage in the UK labour market as a whole.[18]

[15] Rottger et al, 'HATECHECK: Functional Tests for Hate Speech Detection Models' (May 2021), https://arxiv.org/pdf/2012.15606.pdf?mc_cid=90089cdf73&mc_eid=93d4423ac1 accessed 17 June 2021.

[16] The Guardian, 'Police resources 'drained to dangerously low levels', say former top officers' (July 2019), https://www.theguardian.com/uk-news/2019/jul/04/police-watchdog-reforms-chief-inspector-constabulary accessed June 17 2021.

[17] NPCC, 'Dedicated Cybercrime Units Get Million Pound Cash Injection' (April 2019), https://news.npcc.police.uk/releases/dedicated-cybercrime-units-get-million-pound-cash-injection accessed 17 June 2021.

[18] DCMS, 'Cyber Security Skills in the UK Labour Market 2021' (March 2021), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/973802/Ipsos_MORI_Cyber_Skills_in_the_UK_2021_v1.pdf accessed 17 June 2021.

As a result of the lack of resources, criminals are getting off scot free. In 2016-17 the tech industry passed over 70,000 incidents of indecent images of children to the police, Despite this, there were less than 3,500 related convictions.[19] The industry is referring illegal content to the police, but only 5% of cases referred end up being prosecuted.

In addition to greater enforcement of existing laws, education programmes can tackle the root causes of harmful content. Numerous anti-hate programmes being run by the football community with Draw the Line[20] / Against Online Hate[21] campaigns as well as the National Literacy Trust's NewsWise programme for children are good examples of projects that tackle the root cause of problematic content online rather than merely the symptoms.

6. There is already a legal framework to deal with illegal and harmful content

The draft Online Safety Bill ignores the significant and powerful laws that exist to remove existing illegal harms such as child pornography, content that glorifies terrorism, hate speech, even speech deemed 'offensive'. The Public Order Act (1986) covers content identified by civil society groups arguing for tougher laws on harassment and offence. Section 127 of the Malicious Communications Act (2003) covers any communication over any platform that is deemed to be "grossly offensive or of an indecent, obscene or menacing character", with guilt if the perpetrator sends the message for the purpose of "causing annoyance, inconvenience or needless anxiety". Arguments that online trolling or cyber-bullying are outside the existing scope of the law fail to see the strength of the laws already in place.

Beyond this, there are incredibly strict laws on child pornography, on the dissemination of many types of violent pornography, on terrorism, hate speech and incitement to violence.

Proponents of the draft Online Safety Bill have failed to identify specific gaps in the legal framework to deal with contemporary issues arising from the growth of social media.

---

[19] Centre for Policy Studies, 'Safety Without Censorship: A Better Way To Tackle Online Harms' (September 2020), https://www.cps.org.uk/files/reports/original/200926202055-SafetywithoutCensorshipCPSReport.pdf accessed 17 June 2021.

[20] BT, 'Draw The Line: BT launches campaign to tackle online hate' (April 2021), https://www.bt.com/sport/features/draw-the-line-campaign-tackle-online-abuse accessed 17 June 2021.

[21] Sky Sports, 'Against Online Hate: Sky Sports sets out series of measures to fight online hate and abuse' (May 2021), https://www.skysports.com/more-sports/news/29181/12080260/against-online-hate-sky-sports-sets-out-series-of-measures-to-fight-online-hate-and-abuse accessed 17 June 2021.

In the event of such gaps, it should be the role of Parliament, not Ofcom, to decide the parameters of the law in line with the common law and our human rights commitments. Bypassing Parliament deliberately by outsourcing decisions on the parameters of free speech to US and Chinese corporations is a dangerous move away from Parliamentary sovereignty towards private management of our public commons.

Who will be impacted by the draft Online Safety Bill?

The short answer is everyone who uses the internet. The government has repeatedly sought to make out that the draft Online Safety Bill will only impact those who spread hate speech online. That is not the case, the vague and overreaching nature of the Bill means that it will fundamentally impact every single person in the UK who uses the internet for any purpose. Platforms' desire to avoid hefty fines will inevitably lead to over-censorship of innocuous and non-harmful content. A single misconstrued post or overzealous AI can lead to the suspension of your social media accounts, where you communicate with friends and family and store precious memories. While the Bill will impact everyone, certain groups will be particularly impacted:

UNIVERSITIES AND STUDENTS University students will find themselves unable to speak about and debate topics of critical importance online in the same way they would in the lecture hall, which is particularly problematic in the era of remote learning. This is particularly ironic as the draft Online Safety Bill was introduced at the same time as a Bill that aims to protect free speech on campus. As noted by the Open Rights Group's Heather Burns: "an incendiary speaker could deliver a provocative talk in a university lecture hall, and the university would be required to uphold his right to freedom of expression. If he then posts a transcript of his talk onto his social media accounts, those service providers may be required to remove that same speech as being subjectively harmful. The stage is set for a baffling micromanagement of free speech which will make no-one safer."[22]

MARGINALISED PEOPLE As we have already seen the AI employed by tech platforms to implement the provisions of this bill will lead to the silencing of marginalised groups. Many marginalised communities have in recent years reclaimed words that were traditionally used as slurs against them, the inability of AI to distinguish context means these posts will be flagged and removed as hate speech. Victims of trauma or communities struggling with their mental health will find themselves unable to share their own experiences online and gain support from those in similar situations as references to such topics will be considered as potentially harmful.

[22] Heather Burns, ' Why the online safety bill threatens our civil liberties' (May 2021), https://www.politics.co.uk/comment/2021/05/26/why-the-online-safety-bill-threatens-our-civil-liberties/ accessed 17 June 2021.

PEOPLE WHO MAKE JOKES ONLINE The tendency of social media platforms to overcensor coupled with AI's inability to recognise context will mean that millions of ordinary people will find their content taken down and their accounts potentially suspended over remarks that were intended to be humorous or satirical.

PEOPLE LIVING UNDER DICTATORSHIP The tendency of social media platforms to overcensor coupled with AI's inability to recognise context will mean that millions of ordinary people will find their content taken down and their accounts potentially suspended over remarks that were intended to be humorous or satirical.

VICTIMS OF CRIME To prosecute criminals for online wrongdoing , law enforcement agencies require access to the relevant content to use as evidence. If social media companies unilaterally delete illegal content without a proper system for archiving evidence, it will make the job of law enforcement much harder and by extension make all of us less safe.

An Alternative Solution to Online Harms

The Online Safety Bill can be focused, through amendment, to tackle illegal content online. However, the Bill in its current scope is not a viable solution due to the inclusion of the concept of the Duty of Care covering legal content. A report by the Centre for Policy Studies, published in 2020, set out the groundwork for a more targeted and effective model of regulation, in which free speech whether it is offline or online is treated fairly and it is Parliament, rather than regulators, who set the parameters for what is legal and/or harmful and what is not.[23]

The Report proposes:

- A tough new regulator, still under the oversight of Ofcom, but one that should work collaboratively with the police and the Crown Prosecution Service to tackle the criminal activity online through the courts.

- A clearly demarcated regulatory regime for legal vs illegal content. The dividing line offers much greater safeguards against overreach.

[23] Centre for Policy Studies, 'Safety Without Censorship: A Better Way To Tackle Online Harms' (September 2020), https://www.cps.org.uk/files/reports/original/200926202055-SafetywithoutCensorshipCPSReport.pdf accessed 17 June 2021.

- Significant new resources for the police to conduct forensic investigations online and ensure prosecutions can be brought against those publishing illegal content.

- In addition, harms that occur online that are still lawful will be identified and reported to the regulator by stakeholders with relevant expertise and bodies designated to lodge 'Super Complaints' with the regulator.

- The regulator should provide regular thematic reports on the 'Super Complaints' and thought leadership to Parliament who will in turn make recommendations to the Government on additional legislation to address these challenges. This will serve as a vital tool for connecting online harms to democratic scrutiny.

Research undertaken by Heather Burns, Open Rights Group.

1. Content moderation and service design duties

Duty	User to user	Search	All Services	Likely Children*	Category 1	Enforceable by Ofcom	Citation**
Dut Illegal content risk assessment	✓	✓	✓	✓	✓	✓	2 / 2 / 7 2 / 3 / 19
Children's risk assessment	✓	✓	✓	✓	✓	✓	2 / 2 / 7 (9) 2 / 3 / 19 (4)
Adults risk assessment	✓				✓	✓	2 / 2 / 11
Illegal content duties	✓	✓	✓	✓	✓	✓	2 / 2 / 9 2 / 3 / 21
Freedom of expression and privacy duties	✓	✓	✓		✓	✓	2 / 2 / 12
Democratic content duties	✓				✓	✓	2 / 2 / 13
Journalistic content duties	✓				✓	✓	2 / 2 / 14
Reporting and redress duties	✓	✓	✓	✓	✓	✓	2 / 2 / 15
Record keeping and review duties	✓	✓	✓	✓	✓	✓	2 / 2 / 16
Duties to carry out risk assessments	✓	✓	✓	✓	✓	✓	2 / 3 / 17
Safety duties for services likely to be accessed by children	✓	✓		✓		✓	2 / 2 / 10 2 / 3 / 22
Assessments about access by children	✓	✓	✓	✓	✓	✓	2 / 4
Transparency reports	✓	✓			✓	✓	3 / 1 / 49

2. Risk assessment requirements (2 / 2 / 7)

Requirement to identify, assess, and understand:	Illegal content risk assessment	Children's risk assessment	Adults risk assessment
The user base	✓	✓	✓
Risk to users of encountering illegal content (Terror / CSEA)	✓		
Level of harm to users of illegal content	✓		
The number of children accessing the service by age group		✓	
Level of risk to children of encountering each kind of priority primary content		✓	
Each kind of priority primary content that is harmful to children		✓	
Each kind of primary content that is harmful to children or adults, with each one separately assessed		✓	✓
Non-designated content that is harmful to children		✓	
Level of risk of harm presented by different descriptions of content that is harmful for children by age group		✓	✓
Level of risk of functionalities allowing users to search for other members including children		✓	
Level of risk of functionalities allowing users to contact other users including children		✓	
Level of risk to adults of encountering other content that is harmful		✓	
Level of risk of functionalities of the service facilitating the presence or dissemination of illegal content, identifying and assessing those functionalities that present higher levels of risk		✓	
The different ways in which the service is used, and the impact that has on the level of risk of harm that might be suffered by individuals		✓	
Nature, and severity, of the harm that might be suffered by individuals by the above, including children by age group		✓	✓
How the design and operation of the service (including the business model, governance and other systems and processes) may reduce or increase the risks identified.	✓	✓	✓

3. Administrative duties

Duty	User to user	Search	All Services	Likely Children*	Category 1	Enforceable by Ofcom	Citation**
Notify Ofcom for fee payments	✓	✓	✓			✓	3 / 1 / 51
Information notices	✓	✓	✓			✓	4 / 5 / 70
Designation of a person to prepare a report	✓	✓	✓			✓	4 / 5 / 74 (4)
Assist the person preparing the report	✓	✓	✓			✓	4 / 5 / 74 (8)
Cooperate with an Ofcom investigation	✓	✓	✓			✓	4 / 5 / 75 (1)
Attend an interview with Ofcom	✓	✓	✓			✓	4 / 5 / 76 (2)
Duty to make a public statement	✓	✓	✓			✓	6 / 112 (3)
Information in connection with services presenting a threat	✓	✓	✓			✓	6 / 112 (5)

* The draft Bill holds that services which are not using age verification or age assurance to identify the ages of all their visitors will be assumed to be accessed by children. Therefore, "likely to be accessed by children" realistically puts any site or service within a duty of care obligation, regardless of its inclusion or exclusion from that duty of care's legal definition.
** Citation refers to the Bill text. For example, 4 / 5/ 74 (8) means Part 4, Chapter 5, section 74, paragraph 8.

https://www.cps.org.uk/files/reports/original/200926202055-SafetywithoutCensorshipCPSReport.pdf accessed 17 June 2021.