THE STATE OF FREE SPEECH ONLINE

BIG BROTHER WATCH
BigBrotherWatch.org.uk
@BigBrotherWatch

About Big Brother Watch

Big Brother Watch is a civil liberties and privacy campaigning organisation, fighting for a free future. We're determined to reclaim our privacy and defend freedoms at this time of enormous technological change.
We're a fiercely independent, non-partisan and non-profit group who work to roll back the surveillance state and protect rights in parliament, the media or the courts if we have to.
We publish unique investigations and pursue powerful public campaigns.
We work relentlessly to inform, amplify and empower the public voice so we can collectively reclaim our privacy, defend our civil liberties and protect freedoms for the future.

Contact
Silkie Carlo
Director
Email: silkie.carlo@bigbrotherwatch.org.uk
Mark Johnson
Legal & Policy Officer
Email: mark.johnson@bigbrotherwatch.org.uk

Published: September 2021

Contents

Introduction----------**5**

CHAPTER 1: BIG TECH AND THE SHRINKING SPACE FOR FREE SPEECH----------**11**

Private policies vs the law----------**12**

UK communications law----------**12**

Our research----------**13**

Facebook----------**14**

Facebook and hate speech---------- 14

Facebook: Hate speech - sex and gender----------16

Facebook: Hate speech and race----------21

Facebook and COVID-19----------23

"Downplaying severity"----------23

"False", nuanced and complex claims----------24

Facts without context----------26

Fact-checking----------26

Academics----------26

Journalists----------29

Private groups----------29

Facebook and political censorship----------31

Instagram----------**34**

Instagram and hate speech----------34

Instagram: mental Health and self-harm----------36

Guest contribution: It Matters----------**41**

Instagram and COVID-19----------43

Twitter----------**49**

Twitter and hate speech----------49

Men and crime----------50

Misgendering----------53

'Cis' and 'TERF'----------56

Twitter and COVID-19----------61

Twitter and political censorship----------67

YouTube----------**72**
YouTube, hate speech and journalism----------73
YouTube and COVID-19----------75
CHAPTER 2: THE ROLE OF THE STATE----------**81**
Online Safety Bill----------**82**
Duties of care----------83
Potentially illegal content----------84
"Harmful" content----------86
Children----------87
The regulator should not seek to enforce companies' terms and conditions--88
The proposals would erode lawful expression online----------89
Loose definitions of harm----------90
Disinformation----------92
Intimidation----------92
Self-harm----------93
Self-harm: evaluating the case for censorship----------93
An excessive increase in executive power----------95
Harsh punishments will encourage companies' zealous censorship----------96
Technological enforcement----------98
Free expression and privacy duties----------99
Free expression duties----------99
Political and journalistic carveouts----------101
The Bill would further erode the right to privacy online----------103
Private messaging services----------104
Technology notices----------104
The Government's "counter disinformation" activity----------**106**
CHAPTER 3: RECOMMENDATIONS----------**108**
Recommendations for platforms----------**109**
Recommendations for policymakers----------**114**

"Everyone has the right to freedom of opinion and expression; this right includes the freedom to hold opinions without interference and to seek, receive, and impart information and ideas through any media and regardless of frontiers."

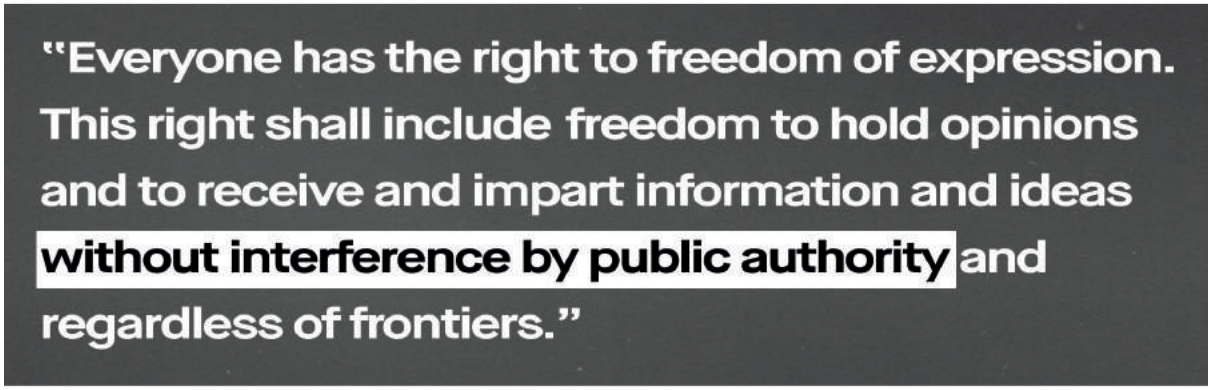
– Article 19,
Universal Declaration of Human Rights

Introduction

Introduction

The right to freedom of speech is the foundation of democracy; the concept is as old as democracy itself. The right to free speech allows people to develop, express and share opinions, ideas and information without undue interference. It is both a progenitor and successor of freedom of thought, self-determination and collective freedom.

International human rights covenants to which the UK is a signatory, including the Universal Declaration of Human Rights, the International Covenant on Civil and Political Rights, and the European Charter of Human Rights (ECHR) impose duties on the UK to ensure an enabling environment for, and to protect, people's right to freedom of expression and information. Article 10 of the Human Rights Act (1998), incorporating ECHR into domestic law, states:



These words are at the heart of the rights movement that followed World War II. The history of democracies has been marked by battles between authoritarian censors and citizens seeking the right to express themselves freely. The ability to speak and publish freely is one of the most empowering for a citizenry – but it is under threat. The technological revolution we are living through presents new opportunities for censorship on a scale never seen before.

The internet can be a democratising force, putting the opportunity of instantaneous global communications at our fingertips. However, these communications are largely controlled by corporate intermediaries. Social media platforms are increasingly becoming our public squares – our modern high streets, meeting rooms and town halls. It can be argued that

social media companies and the information they filter to us are gradually creating the lenses through which we see the world.

Accordingly, social media platforms wield huge economic, social and psychological power. Facebook, for example, has a larger user base than the population of any country in the world,[1] a greater GDP than many nation states,[2] and is even developing its own currency.[3] The company has discovered the power to manipulate moods,[4] sell products,[5] and influence elections.[6]

However, this extraordinary corporate power has never been meaningfully challenged by Western governments, who have made greater efforts at sharing this power than limiting it. Governments are focusing on limiting the speech of their citizens using social media platforms – and the most effective way to do this at scale is not necessarily through law enforcement but to endorse a growing role for the companies as snoopers and speech police. Our current Government seems more concerned that the companies use their power to remove undesirable speech than that they fit into pre-existing structures of due process.

Platforms such as Facebook, Twitter and YouTube have become de facto arbiters of speech online, making judgements about the permissibility of citizens' speech without due process - and not only ordinary citizens, but politicians of the highest office too. This handful of companies has monopolised the modern online communications environment.

Those who argue that online censorship is simply the privilege of a private company often miss the power dynamics truly at play. It must be acknowledged that these are private companies unlike any we have seen before, with power and control over the

1 Facebook has 2.6 billion monthly active users as of March 31, 2020 (https://investor.fb.com/investor-news/press-release-details/2020/Facebook-Reports-First-Quarter-2020-Results/default.aspx). China's population is 1.4 billion. Facebook's annual turnover was $86 billion in 2020 (https://investor.fb.com/investor-news/press-release-details/2021/Facebook-Reports-Fourth-Quarter-and-Full-Year-2020-Results/default.aspx)

2 25 giant companies that are bigger than entire countries – F. Belinchon and R. Moynihan, Business Insider, 25 July 2018 (https://www.businessinsider.com/25-giant-companies-that-earn-more-than-entire-countries-2018-7?r=US&IR=T)

3 Facebook announced the launched of Libra in 2019; the currency, now renamed Diem, is due to launch in 2021.

4 Facebook reveals news feed experiment to control emotions – Robert Booth, the Guardian, 30th June 2014 (https://www.theguardian.com/technology/2014/jun/29/facebook-users-emotions-news-feeds)

5 Success stories – Facebook (https://www.facebook.com/business/success)

6 "Success story: the best content to influence voters" - Facebook (https://www.facebook.com/business/success/toomey-for-senate – last accessed 15.2.21)

speech and privacy rights of citizens comparable to that of governments. The promise that free market competition will liberalise speech is proving to be unrealistic and wrong.

Already in 2021, online censorship has been at the centre of debates about the state of free speech and democracy. In an extraordinary turn of events, former President Donald Trump was suspended and banned from all major platforms and internet services in his last days of office. Other acts of online censorship have taken place within the UK – for example, YouTube removed (and later reinstated) national radio broadcaster talkRADIO's channel, and Labour MP Zarah Sultana's Instagram post about covid deaths was wrongly disallowed for being "false". Each case raises the alarm about social media platforms' interference in free expression.

The Government introduced the draft Online Safety Bill in May 2021 to, in its own words, make "the UK to be the safest place in the world to go online".[7] We agree with Government that the rules that apply in the offline world should apply in the online world – but we do not agree that rules that do not apply in the offline world should be imposed on the online world.

The Bill would see companies tasked with policing the speech of millions, far beyond existing legal boundaries, entrenching and extending privatised monitoring and censorship. Under the Bill, if platforms do not tackle "harmful" content on their site, even though it may be lawful, they may be reprimanded with fines or more severe punishments. This ill-defined approach would result in the state-sanctioned censorship of lawful expression online of a magnitude never previously possible in a liberal democracy.

Under the threat of looming penalties, social media companies will be quicker than ever to enforce their censorious policies. Precisely what types of expression fall into the "legal but harmful" crosshairs remain to be seen. The draft Bill describes such content as that which presents "a material risk of the content having, or indirectly having, a significant adverse physical or psychological impact on an adult of ordinary sensibilities" (Cl. 46(3)). How lawful speech could be policed in this way without eroding freedom of expression is difficult to fathom. We cannot operate policies like this on trust alone - there are clear risks of information control.

7 Online Harms White Paper, - DCMS and the Home Office, April 2019, p.5, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf

The Bill also reflects the national panic, driven by those in positions of authority and power, about "misinformation" and "disinformation". Already, companies' misinformation policies are resulting in censorship that stifle debates and, on numerous occasions, has led to the suppression of information later proved not to be false at all. It is a core principle of post-Enlightenment democracies that an open forum leads to the ongoing discovery of truths. The Government has abandoned this foundational liberal principle. The open forum has been recast as a danger to democracy, where the blunt tool of suppression is preferred to the forces of reason and rationality. In the words of former Supreme Court judge Lord Sumption, "We cannot have truth without accommodating error. It is the price that we pay for allowing knowledge and understanding to develop and human civilisation to advance."

With the rise of automated speech regulation, free expression online is becoming a permitted activity rather than an inalienable right. The permissibility, visibility and legitimacy of speech has been significantly subordinated to corporate authorities who determine whether speech can be expressed, seen, heard, or labelled as false online. Under the Online Safety Bill, these corporate censorship powers are merging with state bureaucracy to exert even greater control over public speech.

The inevitable consequence of the online 'harms' agenda would be the censorship of the most marginalised groups in society. Expression that is unpopular, controversial or counter-cultural will be, if not directly targeted, seen as merely collateral damage under risk-averse, politically-influenced content policies. Not only are the rules oppressive, their enforcement is too. Social media companies rely on algorithms to monitor users' posts and enforce restrictive policies across millions of accounts. In this report, we document how Instagram's algorithm detects self-harm scars in photos, triggering the censorship of people who have experienced mental ill health. This is discriminatory and wrong.

If passed, the Online Safety Bill will set a precedent for the future of free expression online. At present, the draft Bill would severely damage free speech in the UK and set a disastrous precedent for free expression all around the world. The UK Government's aim is to establish a blueprint for international regulation, building "a global coalition of countries all taking co-ordinated steps" on online harms.[8] However, this detachment from the clear boundaries of laws and due process towards nebulous, politicised concepts of "harm" would be a disaster for human rights and freedom of speech in the UK and

8 Online Harms White Paper – DCMS and The Home Office, April 2019, p.6

overseas. This report examines the state of free speech online, mapping the impact of social media companies' corporatisation of speech standards and the Government's role in creating a two-tier speech system. It is the product of over two years of research on online censorship, during which major themes have emerged: "hate speech" including speech on sex, gender and race; political posts, including left-wing and right-wing posts; and posts relating to health, from mental health to Covid-19.

It is fully expected that readers will find some of example banned posts in this report disagreeable, misguided or offensive. However, we ask you not to judge your agreement with these posts, but to probe the more important questions – first, should this lawful content be censored by a private company and second, should lawful speech be censored with the state's backing?

Readers should also note that the examples cited in this report are merely a fraction of the unjustified censorship that we have researched and that, no doubt, has not fallen within the confines of our research. Some readers will, rightly, feel that certain forms of unfair social media censorship are not represented in this report. Our examples are not intended to be a comprehensive or fully representative example – such a task would be impossible, given the scale and opacity of corporate censorship online. However, it is a snapshot of some of the major themes we uncovered in the course of our team's research.

The report goes on to consider the draft Online Safety Bill and why the proposals would materially damage the right to free expression online. Finally, we put forward recommendations for policymakers on how to keep our online space safe and free.

To regulate the internet is to shape the contemporary world and the democratic rights we have within it. The Government's proposals for internet regulation will set norms for new modes of social interaction; inscribe limitations on people's freedom; influence power relationships between businesses, citizens and the state; and write enduring rules into a changing world, for millions of people.

If the UK is to demonstrate leadership on both innovation and human rights, we must transmit the essential principle of free expression and the rule of law to the online realm. This is key for us to construct a modern, digital democracy, built to last.

~~Big Tech and the Shrinking Space for Free Speech~~

CHAPTER 1: BIG TECH AND THE SHRINKING SPACE FOR FREE SPEECH

Private policies vs the law

Online communications are, for the most part, mediated by private profit-making tech companies. Whether for work or personal relationships, we use emails, video calls, messengers, and of course social media platforms to communicate. Politicians and world leaders too, rely on social media platforms to communicate directly to millions of followers around the globe beyond broadcasters, the print press and the traditional media circuit.

However, the major tech companies' rules that govern the online space have become far more restrictive than traditional democratic standards. Driven by brand identity, reputational concerns, and ultimately the bottom line, social media companies' content policies are, in practice, renegotiating access to the free expression rights of democratic populations around the world.

The dominance of Americanised terms of service over modern communications is not just a philosophical problem – it has a real impact on discourse and public freedoms. As the censorship case studies in this paper show, increasingly restrictive terms of service are seriously limiting individuals' ability to speak freely online. This can have serious personal consequences, a wider chilling effect, and political impact too. It is likely that the terms of some modern debates have been seriously skewed by these corporate restrictions on permissible speech. In fact, the dominance of corporate speech controls may be changing conceptions of the right to free speech itself. It is quite possible that these restrictions contribute to the disorientation of public and governmental expectations of permissible expression and censorship.

UK communications law

Communications that are unlawful online should be unlawful online. However, it is important to acknowledge that the UK has particularly extensive legal restrictions on communications, quite unlike the free speech environment of the US. Further, the UK's

broad communications laws do criminalise speech online that is unlawful offline to a similar extent, and this was acknowledged by the Law Commission and by the Government in the 2019 Online Harms White Paper.[9]

The UK already has expansive laws governing speech-related offences that can be used to prosecute violent, hateful and harmful forms of speech and behaviour online. This includes laws prohibiting speech that causes harassment, alarm, distress, or fear (Protection from Harassment Act 1997; Public Order Act 1986); speech that is deemed grossly offensive and purposefully annoying or distressing (Malicious Communications Act 1988; Communications Act 2003); and speech that incites hatred on the basis of race, religion or sexual orientation (Crime and Disorder Act 1998; Race and Religious Hatred Act 2006).

Our research

We analysed the content moderation policies of four major social media platforms through a human rights lens: Facebook, Instagram, Twitter and YouTube. We examined their policies with regard to hate speech, COVID-19 related content and contentious areas such as "misinformation".

Over two years of open-source research, we collected and analysed hundreds of clearly lawful posts that had resulted in significant enforcement – whether they were hidden, removed, or caused the account to be suspended or banned.

Some major themes emerged where enforcement has been questionable, inconsistent and problematic: particularly in relation to gender, sex and race; radical political views; mental health issues; and COVID-19. We explore these areas with examples throughout this chapter.

9 Online Harms White Paper – DCMS and The Home Office, April 2019, p.34

Facebook

Facebook and hate speech

Hate speech is not a specific offence within UK law – rather, a crime that is also motivated by hate or prejudice can be categorised as a 'hate crime'. This is an important distinction.

Facebook introduces its expansive hate speech policy with the justification that such speech is prohibited "because it creates an environment of intimidation and exclusion", not because of any criminal threshold or external legal standards, and further posits that such speech "in some cases, may promote real-world violence"[10]. These suppositions lay the groundwork for a particularly strict environment of speech enforcement.

Facebook defines hate speech as "a direct attack" on certain identified groups, meaning "violent or dehumanising speech, harmful stereotypes, statements of inferiority or calls for exclusion or segregation."[11] Facebook expands on the meaning of "attack" giving examples such as referring to an identified group or group member in the context of "mocking the concept, events or victims of hate crimes", making "generalisations or unqualified behavioural statements (in written or visual form)", "animal", "insect" or "filth" comparisons, or references to members as "criminals" or "sexual predators".[12] Whilst speech that causes a threat of violence is prohibited in UK law, speech that suggests inferiority, exclusion or segregation, or that is dehumanising, would not automatically be illegal.

The policy also describes an expansive approach to "generalisations that state inferiority" which can include any expression that another group or group member is "less than adequate", or better or worse than another group.[13] Perhaps the most poorly defined prohibition in this policy section is expression about "deviating from the norm", in which "the norm" is not defined. The policy specifically warns that this makes words such as "useless" and "abnormal" problematic in this context.[14]

10 Facebook, Objectionable Content, Community, Guidelines, https://www.facebook.com/communitystandards/objectionable_content
11 Ibid.
12 Facebook, Hate Speech, Objectionable Content, Community Guidelines, https://www.facebook.com/communitystandards/hate_speech/
13 Ibid.
14 Ibid.

The policy also specifies the inclusion of expressions of inferiority that reference hygiene, smell, physical appearance, intellectual capacity and education (specifically including the words "stupid", "idiots", "uneducated"), mental health (specifically including "crazy", "insane"), moral deficiencies ("arrogant", "coward"), and derogatory sexual terms ("perverts", "slut").[15]

"Contempt" is specifically prohibited, including homophobia and Islamophobia but extending further to expressions of dismissal such as "don't like, don't care for" a group or group member, expressions of repulsion (the word "yuck" is specified), and cursing (the phrase "kiss my ass" is specified).[16]

In a further critical divergence from existing legal standards, Facebook applies these rules to groups that fall within "what we [Facebook] call protected characteristics". Those protected characteristics are broad, covering not only the five protected groups of disability, race (and national origin[17]), religion, sexual orientation and gender reassignment as per the UK's hate crime definition, but also sex, gender, caste and serious disease.[18] The platform extends this further to include immigration status for the types of "attack" described above. However, immigration status is not one of Facebook's protected characteristics for prohibited content concerning segregation, political, economic or social exclusion.

The prohibition on expressions of social exclusion specifically includes denial of opportunity to gain "access to spaces (incl. Online)" to any of Facebook's protected groups.[19] This conflicts with some social norms whereby we have, for example, single sex spaces and services, or groups exclusive to individuals with a certain protected characteristic, protected under the Equality Act 2010.[20] According to the community guidelines, calling for exclusion of men from an online space, and vice versa, could be

15 Facebook, Hate Speech, Objectionable Content, Community Guidelines, https://www.facebook.com/communitystandards/hate_speech/
16 Ibid.
17 The definition of race in the Equality Act 2010 includes colour, nationality, and ethnic or national origins (s. 9(1)).
18 Facebook, Hate Speech, Objectionable Content, Community, Guidelines, https://www.facebook.com/communitystandards/hate_speech/
19 Ibid.
20 Citizens Advice UK, Discrimination in the provision of goods and services, https://www.citizensadvice.org.uk/consumer/discrimination-in-the-provision-of-goods-and-services/discrimination-in-the-provision-of-goods-and-services1/goods-and-services-what-are-the-different-types-of-discrimination/what-doesn-t-count-as-unlawful-discrimination-in-goods-and-services/single-sex-and-separate-services-for-men-and-women-when-discrimination-is-allowed/

considered hate speech and duly prohibited.

Facebook: Hate speech - sex and gender

Facebook's definitions of hate speech mean that statements such as "men are scum" or "men are pigs" are prohibited on Facebook, although not prohibited under UK or US law.

This issue came to the fore for Facebook in the wake of the #MeToo movement in Autumn 2017, when comic Marcia Belsky was banned from the platform for 30 days for writing "men are scum" on a woman's post about misogyny she had experienced.[21] A tide of subsequent posts by women echoing the language were removed and enforcement action was taken against the posters.

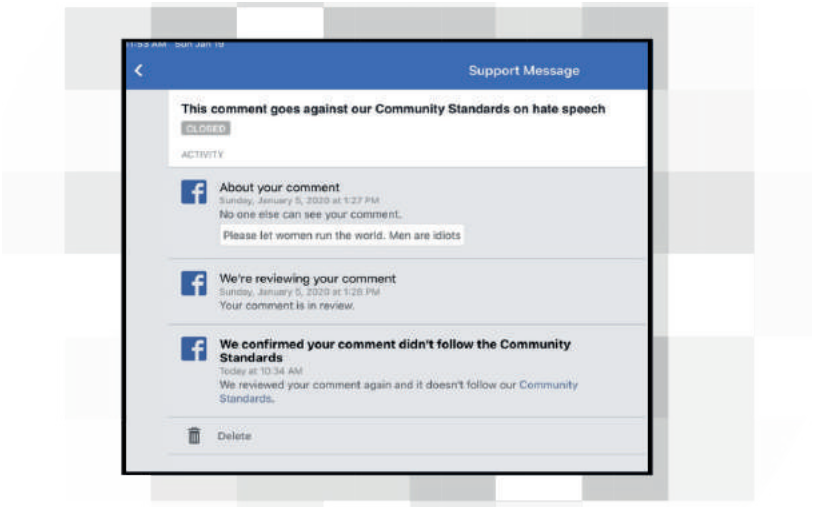
Comedian Maria Belsky is temporarily blocked from posting on Facebook after commenting 'Men are scum' on a woman's post about male sexual assault.

21 Marcia Belsky, Twitter, https://twitter.com/MarciaBelsky/status/921082758574854146

In the months after this high-profile enforcement, Facebook's senior staff - including CEO Mark Zuckerberg - agonised over what the right content policies around this language should be in a prolonged string of senior meetings. Staff floated suggestions such as of treating disqualified content about gender less harshly than that concerning race, or developing different rules for speech about men than women.[22] The policy was not eventually changed and still remains in place.

This policy means that Facebook users continue to be censored and punished for light-hearted and lawful comments.

In another example, a user's comment breached Facebook's Community Standards on hate speech: 'Please let women run the world. Men are idiots'.

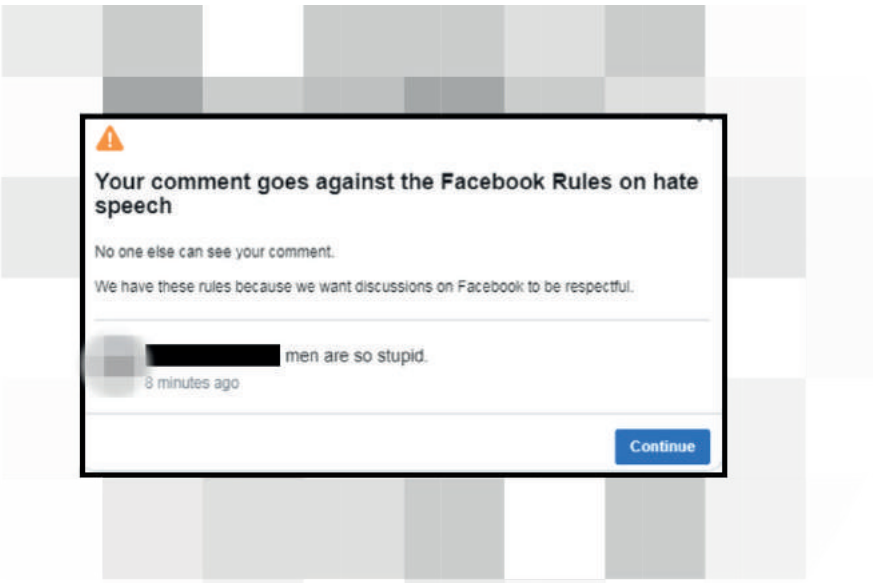


A Facebook user's comment that 'men are idiots' is classed as hate speech by Facebook and consequently blocked from public view.

The policy itself is excessive and its enforcement frequently lacks any contextual consideration.

22 Simon Van Zuylen-Wood ,"Men Are Scum": Inside Facebook's War on Hate Speech , Vanity Fair, 2019: https://www.vanityfair.com/news/2019/02/men-are-scum-inside-facebook-war-on-hate-speech

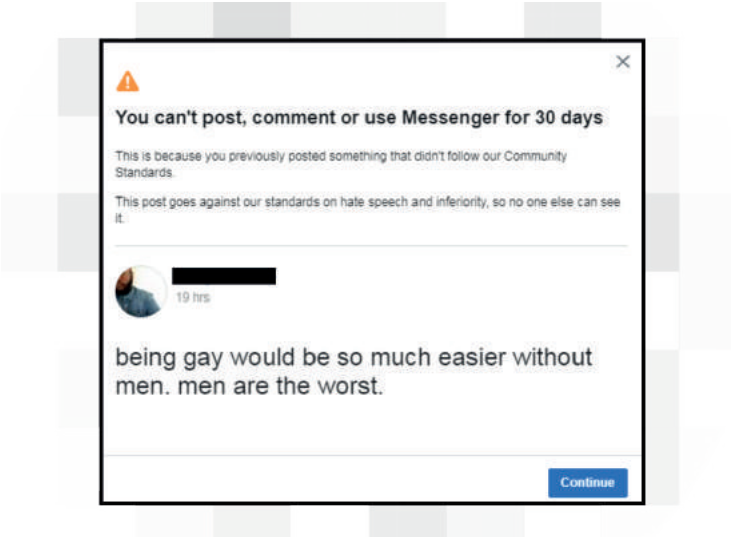
In the example below, a man is censored for "hate speech" after commenting, "men are so stupid".



A Facebook user (a man) comments that 'men are so stupid' and breaches hate speech rules on Facebook. The comment is consequently blocked from public view.

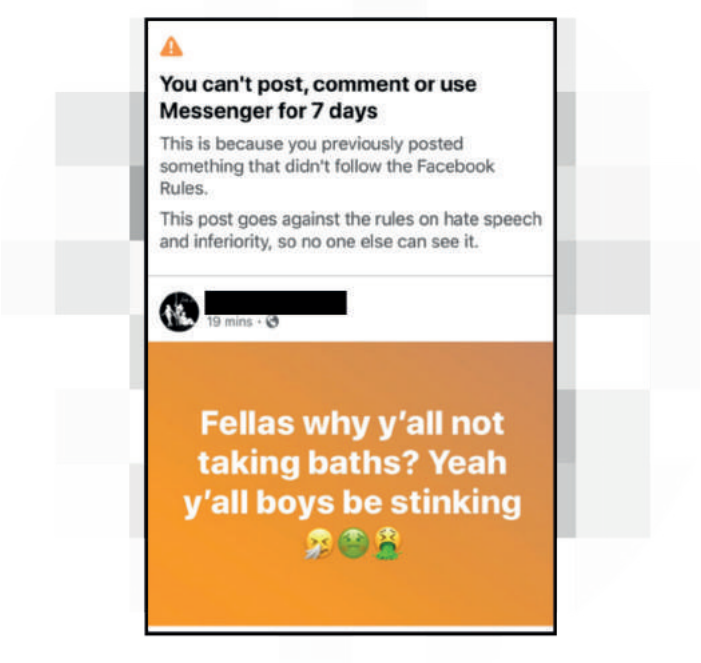
Facebook applies these standards to Messenger too – a widely-used, Facebook-owned instant messenger for text and video that does not require a Facebook profile and can be used independently of the site.

In this example, a man jokes that being gay "would be so much easier without men" because "men are the worst". This falls foul of Facebook's hate speech policy on inferiority and gender. Consequently, the comment was hidden from public view and the user received a 30-day ban from Messenger.



A man using Messenger, owned by Facebook, jokes that "men are the worst" resulting in a 30-day ban.

Another male user was banned from Messenger for 7 days after joking that boys stink. Whilst this will seem absurd to many, it is an accurate application of Facebook's hate speech policy.

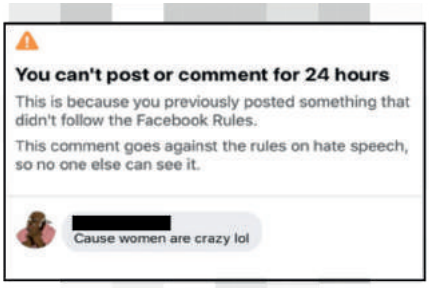


Bringing Facebook's hate speech policies on inferiority (smell and hygiene) to life, this Facebook user (a man) was blocked from Messenger for a week for a light-hearted comment.

The hate speech and gender policy is also applied to comments about women.

Women in particular can often become targets of online abuse, some of which passes a criminal threshold. Serious abuse of women online – particularly stalking, death threats, rape threats and violence - should be taken more seriously by social media platforms and, critically, by the police who should be involved in any such reports.

However, the bar set in Facebook's policy is too low and enforcement too erratic.

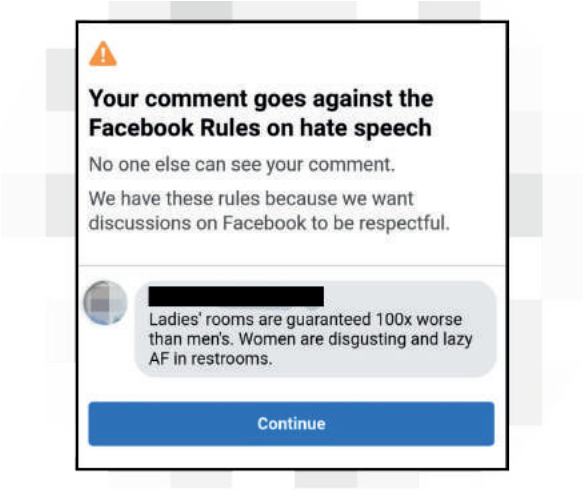


Man blocked on Facebook for joking "women are crazy"

In this example, a man was blocked from using Facebook for one day after breaching the rules on hate speech for apparently joking, "women are crazy lol".

Again, enforcement of the expansive policy lacks consideration of context or nuance.

This woman's comments on ladies' restrooms was censored as it falls foul of Facebook's hate speech (hygiene, inferiority) and gender policy.



This woman's comment on Facebook was censored for suggesting ladies' restrooms are worse than men's because women are "disgusting and lazy AF [as fuck] in restrooms", thus falling foul of the hate speech (inferiority, hygiene) and gender policy.

Facebook's content moderation was the source of some public ridicule in early 2021 after the platform was forced to apologise for removing content regarding "Plymouth Hoe", an area located within the English city of Plymouth. Facebook apologised for mistaking the name for a derogatory misogynistic term.[23]

This widely-reported case demonstrates the blunt nature of platforms' content moderation systems which fail to account for context or nuance and often use algorithms to police overly-broad rules which govern otherwise lawful speech.

23 BBC, Facebook apologises for Plymouth Hoe 'error', 27 January 2021: https://www.bbc.co.uk/news/uk-england-devon-55827981

Facebook: Hate speech and race

Facebook's hate speech policy, which goes far beyond the limits of speech offences in UK law, prohibits "generalisations" in relation to race.[24] This basic and broad policy does not require an assertion of inferiority, superiority or hostility,

Whilst well-intended, the policy seems not to have been fully thought through. The result is that the policy is censorious and stifles open discussion about race and racism.

Facebook users have developed online slang to avoid automated flagging, such as "wypipo" to refer to white people. Some black Facebook users refer to being "Zucked" to describe being suspended after talking about their experiences or analysis of racism.[25]

A teacher and activist in the US, Carolyn Wysinger, shared a post on Facebook about an interview with Hollywood actor Liam Neeson. In the interview, in which Neeson is promoting a film about revenge, he claimed that after learning a friend had been raped by a black man, he had spent some time "hoping I'd be approached (...) by some 'black bastard' (...) so that I could kill him".[26] Wysinger shared the post with her friends, commenting, "White men are so fragile and the mere presence of a black person challenges every single thing in them." Wysinger reported that Facebook deleted the post within fifteen minutes and warned that if she reposted it, she would be banned for 72 hours. Wysinger described the social media platforms' interventionism in race discussions as "exhausting" and emotionally draining.[27]

This enforcement upheld Facebook's broad policy against race-based generalisations – but it also repressed dialogue among friends about their experience of racism. Facebook's expansive and rigid policy creates barriers to discourse in the digital space that are out of step with wider society. Books and articles have been written about "white fragility"[28] but the term is not always permitted in context on Facebook. Natasha Marin, an artist and anti-racism consultant, remarked "Black people are punished on Facebook for speaking

24 Facebook, Hate Speech, Objectionable Content, Community Guidelines, https://www.facebook.com/communitystandards/hate_speech/

25 Guynn, J. Facebook while black: Users call it getting 'Zucked,' say talking about racism is censored as hate speech, USA Today, https://eu.usatoday.com/story/news/2019/04/24/facebook-while-black-zucked-users-say-they-get-blocked-racism-discussion/2859593002/

26 Liam Neeson in racism storm after admitting he wanted to kill a black man, BBC News, 2019, https://www.bbc.co.uk/news/entertainment-arts-47117177

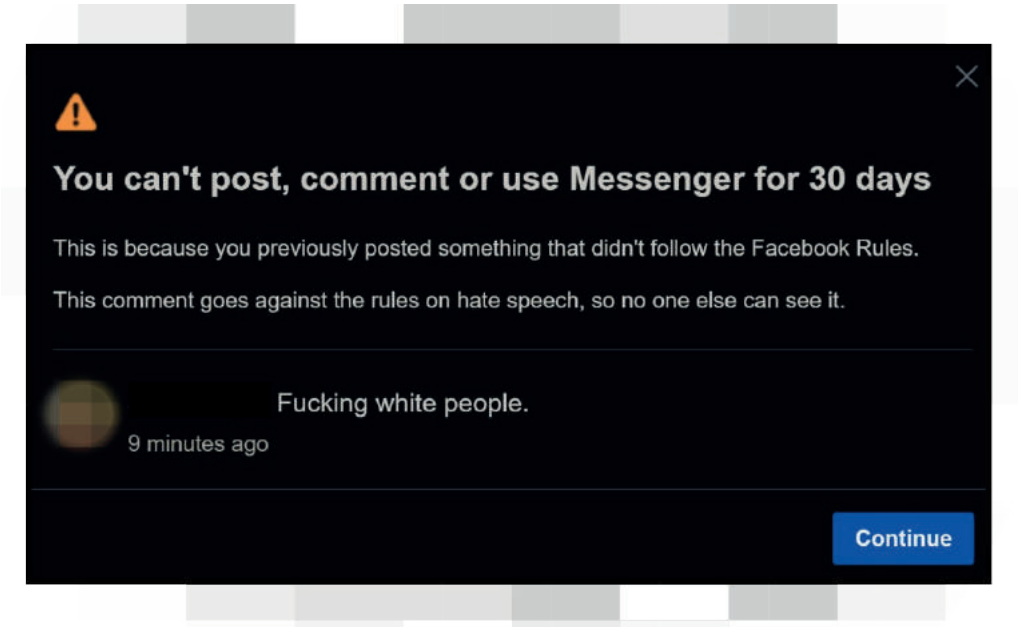
27 Guynn, J. Facebook while black: Users call it getting 'Zucked,' say talking about racism is censored as hate speech, USA Today, https://eu.usatoday.com/story/news/2019/04/24/facebook-while-black-zucked-users-say-they-get-blocked-racism-discussion/2859593002/

28 See DiAngelo, R. White Fragility, June 2018

directly to the racism we have experienced."[29]

As demonstrated by many of the examples here, Facebook's content moderation and policies make no room for an appreciation of context, meaning that freedom of expression standards can never be met.

In the following examples, white Facebook users were censored and suspended under Facebook's hate speech policy for off-hand and even jovial comments about white people.



In this example, a white man receives a 30-day ban from Facebook for "hate speech" against white people.

29 Guynn, J. Facebook while black: Users call it getting 'Zucked,' say talking about racism is censored as hate speech, USA Today, https://eu.usatoday.com/story/news/2019/04/24/facebook-while-black-zucked-users-say-they-get-blocked-racism-discussion/2859593002/

A white woman received a 3-day suspension from Facebook for joking that "every white cop" looks like a cartoon shark. This joke breached the hate speech policy on race.

White woman suspended on Facebook for "racist" shark joke

Facebook and COVID-19

Like many major online platforms, Facebook introduced strict new content moderation rules during the COVID-19 pandemic. These new community guidelines are largely borne out of fears regarding the spread of "misinformation" online regarding both the virus and vaccines.

"Downplaying severity"

Facebook's community guidelines on "Violence and incitement" state that "Misinformation and unverifiable rumours that contribute to the risk of imminent violence or physical harm"

are not permitted on the site.[30] According to Facebook, this policy has been in place since 2018, but since January 2019 the platform has "applied this policy to misinformation about COVID-19."[31] In applying this policy to COVID, content such as "claims that downplay the severity of COVID-19" is prohibited, for example, "claims that the number of deaths caused by COVID-19 are much lower than the official figure".[32] However, the Telegraph reported in April 2021 that in some weeks a quarter of the reported COVID deaths were not caused by the virus but rather represented people who died from other causes whilst infected with the virus.[33] This story is evidence, if needed, that is important to maintain open debate on issues of statistical reporting.

"False", nuanced and complex claims

For content which was deemed to be misleading but not harmful, Facebook had previously applied fact-checking services and labelling. However, on 8th February 2021, the platform expanded what it described as "the list of false claims we will remove to include additional debunked claims about the coronavirus and vaccines."[34] This list includes a number of claims which appear to create no risk of "harm" to users, as we explore in this section.

One of those allegedly misleading claims was that "COVID-19 is man-made or manufactured"[35] despite evidence of the origins being inconclusive. However, after it emerged the US administration was investigating the "lab leak" theory,[36] despite months of censoring posts that explored this possibility, Facebook said it would reverse the ban on 26th May 2021. A Facebook spokesman said,

30 Facebook Community Guidelines, Violence and incitement, Violence and criminal behaviour, https://www.facebook.com/communitystandards/credible_violence

31 Facebook, Combating COVID-19 Misinformation Across Our Apps, 2020, https://about.fb.com/news/2020/03/combating-covid-19-misinformation/

32 Facebook, COVID-19 policy updates and protections: https://www.facebook.com/help/230764881494641 [accessed 17 May 2021]

33 Sarah Knapton and Ben Riley-Smith, The Telegraph, Quarter of Covid deaths not caused by virus, new figures show – 13 April 2021: https://www.telegraph.co.uk/news/2021/04/13/quarter-covid-deaths-not-caused-virus/

34 Facebook, An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19, https://about.fb.com/news/2020/04/covid-19-misinfo-update/#removing-more-false-claims

35 Ibid.

36 Covid: Biden orders investigation into virus origin as lab leak theory debated – BBC, 27th May 2021: https://www.bbc.co.uk/news/world-us-canada-57260009

> "In light of ongoing investigations into the origin of Covid-19 and in consultation with public health experts, we will no longer remove the claim that Covid-19 is man-made from our apps. We're continuing to work with health experts to keep pace with the evolving nature of the pandemic and regularly update our policies as new facts and trends emerge."[37]

Facebook's policy also prohibits claims on issues that are nuanced or complex, and thus could lead to debates being stymied or even posts or accounts being removed despite content being factual. The policy prohibits the claim that COVID vaccines "were developed, produced or designed from/with, human tissue from aborted foetuses/aborted foetal tissue"[38] - however, the development of the AstraZeneca vaccine involved propagating the virus using cell lines that derive from an aborted foetus,[39] as is the case for several other vaccines. The policy prohibits the claim that "COVID-19 was predicted, including in Event 201's pandemic exercise in October 2019"[40] which risks leading to arbitrary or semantic censorship, since Event 201 in October 2019 did not predict but "simulate(d) an outbreak of a novel zoonotic coronavirus transmitted from bats to pigs to people that eventually becomes efficiently transmissible from person to person, leading to a severe pandemic."[41] Further, Facebook's policy also prohibits the claim that COVID vaccines can "seriously harm people (such as causing blood clots)",[42] despite the fact that multiple countries, including the UK, have altered their vaccination programmes in light of the risk, albeit rare, of dangerous blood clots associated with the AstraZeneca vaccine.[43]

37 Facebook lifts ban on posts claiming Covid-19 was man-made – Alex Hern, the Guardian, 27th May 2021: https://www.theguardian.com/technology/2021/may/27/facebook-lifts-ban-on-posts-claiming-covid-19-was-man-made

38 COVID-19 policy updates and protections – Facebook: https://www.facebook.com/help/230764881494641 (accessed 18th May 2021)

39 There are no foetal cells in the AstraZeneca Covid-19 vaccine – Grace Rahman, Full Fact, 26 November 2020: https://fullfact.org/online/foetal-cells-covid-vaccine/

40 COVID-19 policy updates and protections – Facebook: https://www.facebook.com/help/230764881494641 (accessed 18th May 2021)

41 The Event 201 Scenario – Event 201: https://www.centerforhealthsecurity.org/event201/scenario.html (accessed 18 May 2021)

42 COVID-19 policy updates and protections – Facebook: https://www.facebook.com/help/230764881494641 (accessed 18th May 2021)

43 COVID-19 vaccination and blood clotting – Public Health England, 7th May 2021: https://www.gov.uk/government/publications/covid-19-vaccination-and-blood-clotting/covid-19-vaccination-and-blood-clotting

Facts without context

The policy even implies certain editorial expectations of users, stating that accounts may be removed for sharing "shocking stories" related to COVID-19, explicitly stating this may apply to "actually true events or facts that raise safety concerns", if they are presented "without context".[44] Such restrictive policies are unlikely to boost trust among low-trust communities where it matters most, but to reinforce concerns about power and the erosion of free speech.

Fact-checking

Facebook's policy states that even where COVID or vaccine-related posts do not violate these expansive policies, posts "will still be eligible for review by our third-party fact-checkers, and if they are rated false, they will be labelled and demoted."[45] Describing how this process works, Facebook states that "Once a post is rated false by a fact-checker, we reduce its distribution so fewer people see it, and we show strong warning labels and notifications to people who still come across it, try to share it or already have."[46] The platform set out that they "regularly update the claims that we remove based on guidance from the WHO and other health authorities."[47] Consequently, these policies have resulted in Facebook adjudicating on points of fact. This move, which editorialises the platforms' content moderation processes, has proved damaging to free expression on the site. The desirability and suitability of a Silicon Valley-based tech platform to act as a fact-checker on a fast-moving public health crisis has also come into question.

Academics

Facebook's fact-checking service directly inhibited discussion around public health policy on the site, labelling an article examining the efficacy of masks by the Oxford

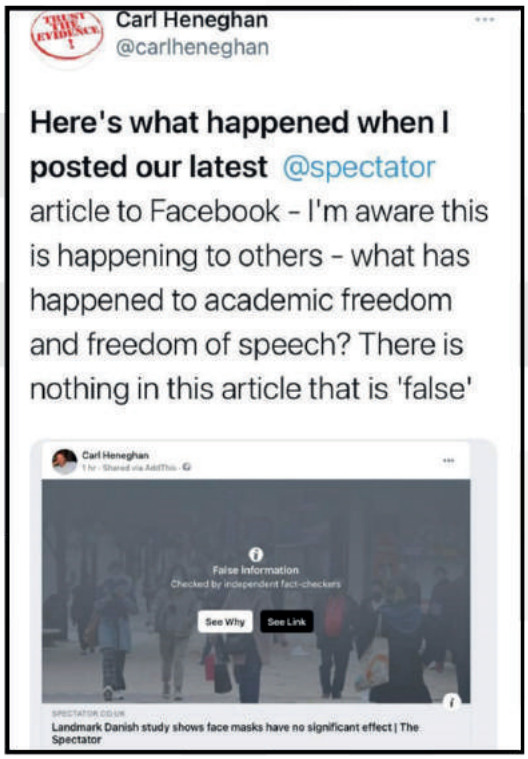
44 COVID-19 policy updates and protections – Facebook: https://www.facebook.com/help/230764881494641 (accessed 18th May 2021)

45 An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19 – Guy Rosen, Facebook, 16th April 2020: https://about.fb.com/news/2020/04/covid-19-misinfo-update/

46 Facebook, Combating COVID-19 Misinformation Across Our Apps, 2020, https://about.fb.com/news/2020/03/combating-covid-19-misinformation/

47 Ibid.

University Professor Carl Heneghan and published in the Spectator in November 2020[48] as "false information", having been "checked by independent fact-checkers".



Oxford University Professor Carl Heneghan took to Twitter to share a screenshot of his article branded as "false information" by Facebook

The disputation of academics' work demonstrates how oppressive Facebook's terms of use have become. The editor of the British Medical Journal Kamran Abbasi, whilst disagreeing with Professor Heneghan's analysis, criticised Facebook's action, writing that "disagreement among experts, especially about interpretation of a study, is a common occurrence. It is the usual business of science."[49] Abbasi wrote:

> "It seems 2020 is Orwell's 1984, where the boundaries of public discourse are governed by multibillion dollar corporations (in place of a totalitarian regime) and secret algorithms coded by unidentified employees. (...) Facebook in particular purports to allow freedom of speech on its platform but acts s e l e c t i v e l y, seemingly without logic, consistency, or transparency. That is how c o n t r o l

48 Landmark Danish study finds no significant effect for facemask wearers – Carl Heneghan and Tom Jefferson, The Spectator, 19th November 2020: https://www.spectator.co.uk/article/do-masks-stop-the-spread-of-covid-19-

49 The curious case of the Danish mask study – Kamran Abbasi, BMJ, 26th November 2020: https://www.bmj.com/content/371/bmj.m4586

of facts and opinions furthers hidden agendas and manipulates the public."[50]

Facebook's intervention was a clear interference with free expression and academic enquiry, and discredited the company's insistence that content moderation does not stifle debate on the platform. Neither public discourse nor scientific debate should be inhibited by fear-induced censorship. It is more important than ever, in the midst of a crisis, that free and open discussions on matters of emerging science and public policy can take place.

This was a case raised by the House of Lords Communications and Digital Committee during a session of the Committee's inquiry into freedom of expression online with executives from Facebook and Twitter. During the session, Chair of the Committee, Lord Gilbert asked a representative from Facebook about the company's decision to label the article, written by Carl Heneghan, as "false information".

Lord Gilbert asked "how qualified an expert the fact-checker would have been to analyse all of this other information and the view of other experts, in order to come to a view that what Carl Heneghan, Professor of Evidence-based medicine at Oxford, had said was false?"[51]. He also asked if they would have had any medical or scientific qualifications.[52]

Lord Gilbert said that fact checkers themselves profess their role to "provide context for people to make up their own minds" and disputed whether Facebook's "false information" label, actually did this in practice.[53] During the session the Committee Chair also took the opportunity to ask a representative from Twitter about the platform's decision to mark a tweet by a Professor of Medicine at Harvard Medical School, about vaccines, as misleading.[54]

Public discourse should not be inhibited by fear-induced censorship. It is more important than ever that free and open discussions on matters of public policy can take

50 The curious case of the Danish mask study – Kamran Abbasi, BMJ, 26th November 2020: https://www.bmj.com/content/371/bmj.m4586

51 House of Lords Communications and Digital Committee, Parliament TV, 27 April 2021, https://parliamentlive.tv/event/index/cdcfadb9-6594-4e3a-99d5-37645b55c935

52 Ibid.

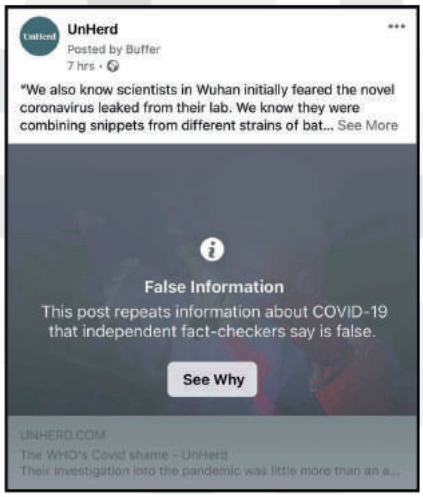
53 Ibid.

54 Ibid.

place in the face of a crisis.

Journalists

Similarly, the award-winning investigative reporter Ian Birrell was subjected to Covid "fact-checking" and labelling on Facebook. In February 2021, Birrell's article in UnHerd, which criticised the WHO's investigation into the origins of COVID-19, was flagged on Facebook as "false information".[55]



Award-winning reporter Ian Birrell's article on Unherd, criticising the WHO's investigation of the origins of COVID-19, was marked as "false information" on Facebook

It was right for Birrell to pursue lines of enquiry and scrutinise the claims of powerful institutions, as should any journalist. The idea that Facebook should repress journalistic scrutiny of an international health organisation's work is chilling and actively anti-democratic.

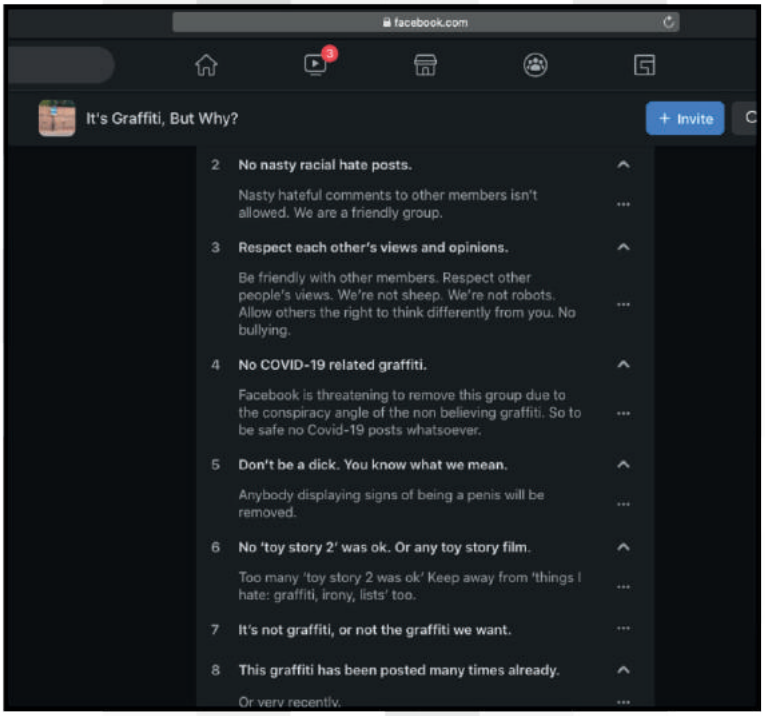
Private groups

Facebook's policies on COVID-19 have also been applied in private groups using the platform, even if members depict others' views on COVID, if those views breach these broad policies. Members of a graffiti image-sharing group called "It's Graffiti, But Why?"

55 Facebook censors award-winning journalist for criticising the WHO – Freddie Sayers, Unherd, 11th February 2021: https://unherd.com/thepost/facebook-censors-award-winning-journalist-for-criticising-the-who/

were warned that the page would be closed down by the platform if members continued to post graffiti images about COVID-19 (and the page was indeed later removed).[56] The group warned, "Facebook is threatening to close us down due to the conspiracy angle of the non believing graffiti"(see screenshot). The group administrator urged members not to post content relating to the pandemic after Facebook threatened to close it down.

It is not clear whether posts that simply documented the existence of graffiti which contained COVID-19 conspiracy theories were enough to constitute this threat.



The now deleted graffiti image-sharing group, "It's Graffiti, But Why" warned that Facebook had threatened to remove the group for COVID-19 related graffiti images – Facebook did later remove the group.

56 Where did 'it's graffiti, but why?' go? - Private group, Facebook: https://www.facebook.com/groups/141819904330852 (accessed 20th May 2021)

Facebook and political censorship

Facebook's content moderation processes can have particularly negative effects where it affects political speech, which is typically afforded heightened protections in human rights jurisdictions.

On 22nd January 2021, Facebook removed the left-wing Socialist Workers Party from its site. According to an official statement, Facebook executives later "restored the Socialist Workers party Facebook Page and several accounts after an automation error" which the company apologised for.[57]

Whilst social media platforms have become a key tool in the armoury of political campaigners, censorship of political activists' accounts and content on those platforms appears to be increasing. Facebook has previously stated its intention to depoliticise the site, with one Facebook spokesperson arguing that "some people feel that there's too much political content in their news feeds."[58] However, interventions with political speech can easily result in a lack of diversity, suppression of minority or alternative views, and political censorship.

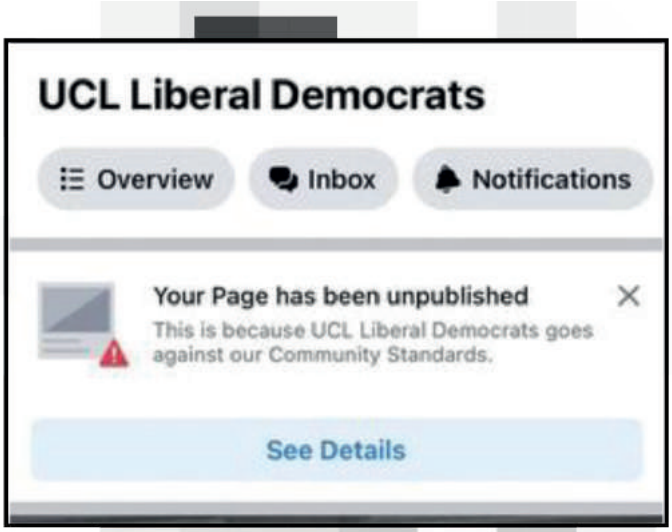
For Facebook to remove an Electoral Commission-registered political party from its platform was a worrying development and a violation of the principles of free expression and wider democratic participation.

Facebook stated that the removal of the Socialist Workers Party's Page from the platform had been an "automation error."

57 Facebook sparks anger after shutting socialist pages, Financial Times, 2020, https://www.ft.com/content/7364679e-dc4f-41fb-9bc9-3eb170159673

58 Lima, C. and Schneider, E. Zuckerberg's pledge to depoliticize Facebook hits grassroots movements, Politico, 2020, https://www.politico.com/news/2021/01/29/facebook-political-groups-grassroots-organizers-463922

This kind of arbitrary account suspension was also dealt to a university branch of the Liberal Democrat Party. The UCL Liberal Democrats found that their account had been suspended by Facebook with a short notification message, which failed to explain why the platform had shut down the account.



The account was later reinstated and Facebook sent the university group the following message, which gave neither an explanation nor apology for the account suspension:

"After reviewing your appeal, your Page UCL Liberal Democrats has been published. This means it can now be viewed publicly."

The arbitrary removal of a party-political account in the midst of a local election campaign is bad for democracy and sets a dangerous precedent for the possible removal of political accounts during the fast pace of prominent election campaigns in the future. In such instances, high-profile takedowns of this nature could genuinely affect election results and undermine the democratic process.

"With the rise of [redacted] automated speech regulation, free expression online is becoming [redacted] a permitted activity [redacted] rather than an inalienable right. [redacted] Under the [redacted] Online Safety Bill, corporate censorship powers [redacted] are merging with state bureaucracy to [redacted] exert even greater [redacted] control over public speech."

BIG BROTHER WATCH

Instagram

Instagram and hate speech

Facebook's ownership of the photo-sharing platform Instagram means that many of the community guidelines on Instagram are the same as those on the site of its parent company. Similarly to Facebook, the "protected characteristics" recognised by the site are race (ethnicity and national origin), disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease.[59]

Like Facebook, "hate speech" against people within these groups constitutes "a direct attack" against the individual(s) in question, which could be "violent or dehumanising speech, harmful stereotypes, statements of inferiority or calls for exclusion or segregation."[60]

Instagram was quick to employ automated content moderation systems, using machine learning from 2018 to detect content that could be illegal or contravene the site's terms of use.[61] However, content moderation of this kind is often a blunt and imprecise tool, which fails to account for context or nuance. As such, AI-powered content moderation systems often censor speech that is both legal and compliant with platforms' terms of use.

One such area in which these tools have struggled has been in-group linguistic reappropriation. This is where a group reclaims certain pejorative words or phrases that were previously directed at the group as terms of abuse. Minority or marginalised groups can source empowerment from the reappropriation of terms in a way that nullifies their previous meaning. This is also common in comedy, where a comedian from a minority background might acceptably satirise otherwise offensive terms.

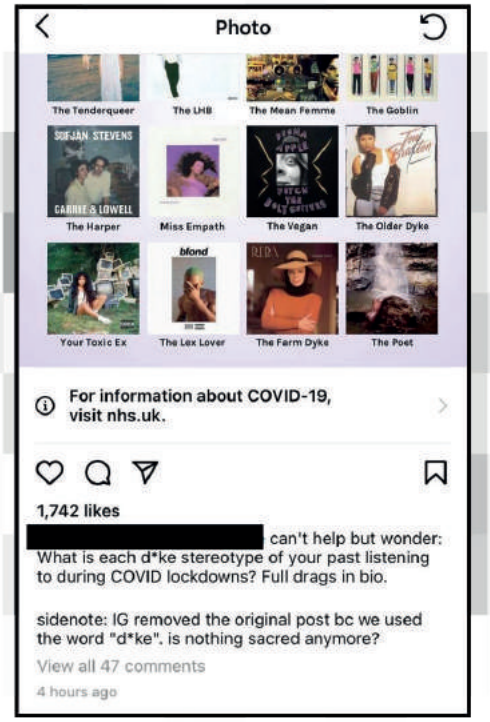
In the example below, an LGBT Instagram page was forced to reissue a post which had

59 Community Guidelines, Instagram, https://www.facebook.com/help/instagram/477434105621119/

60 Facebook, Objectionable Content, Community, Guidelines, https://www.facebook.com/communitystandards/objectionable_content

61 Vincent, J. Instagram is using AI to detect bullying in photos and captions, The Verge, 2018, https://www.theverge.com/2018/10/9/17954658/instagram-ai-machine-learning-detect-filter-bullying-comments-captions-photos

previously contained the word "dyke" in the caption. The term, historically used as a homophobic slur to describe lesbians, has been reclaimed by the LGBT community and can be used jovially, satirically or even as a term of empowerment.



Instagram removed a post from an LGBT+ page as the caption contained the word "dyke".

Examples like this demonstrate how, rather than help marginalised groups, algorithmic content moderation systems can cause disproportionate censorship due to automated systems' inability to detect context or nuance.

This inadvertent censorship or marginalised groups can be found across various types content on the platform – notably, in relation to mental health. Instagram's Community Guidelines state:

> "The Instagram community cares for each other, and is often a place where people facing difficult issues such as eating disorders, cutting, or other kinds of self-injury come together to create awareness or find support. We try to do our part by providing education in the app and adding information in the (Instagram) Help Center so people can get the help they need."[62]

However, numerous counts of censorship have been reported by users of the platform, particularly affecting the accounts of those who have experienced self-harm.

According to the rules on the site:

> "Encouraging or urging people to embrace self-injury is counter to this environment of support, and we'll remove it or disable accounts if it's reported to us. We may also remove content identifying victims or survivors of self-injury if the content targets them for attack or humor."[63]

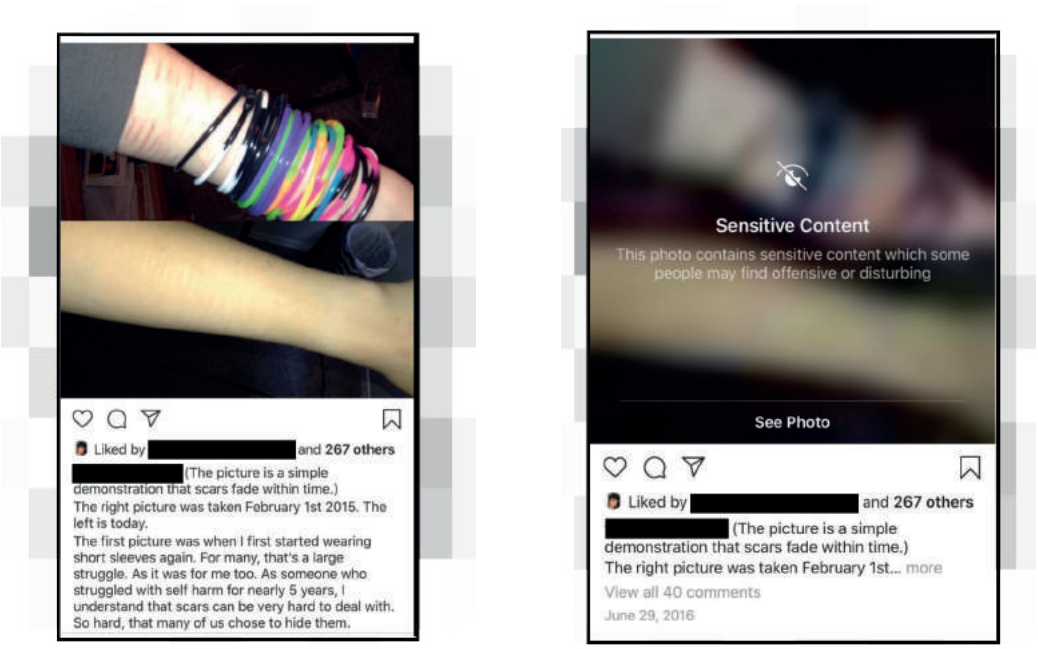
What constitutes "encouraging" self-harm is not defined and appears to be excessive in its application. The result is that "existing" after self-harm can seem to be a violation of Instagram's terms in and of itself. Many posts legitimately documenting content relating to this subject area in any way, including recovery, empowerment and healing, are removed. Furthermore, some posts that are not about self-harm at all but that contain a photo of an individual with scars have been obscured and marked as "sensitive."

Upon stating that the company will remove content that identifies victims or survivors of self-harm for "attack or humor", Instagram appears not to rule out removing the content of users who are victims or survivors of self-harm themselves.

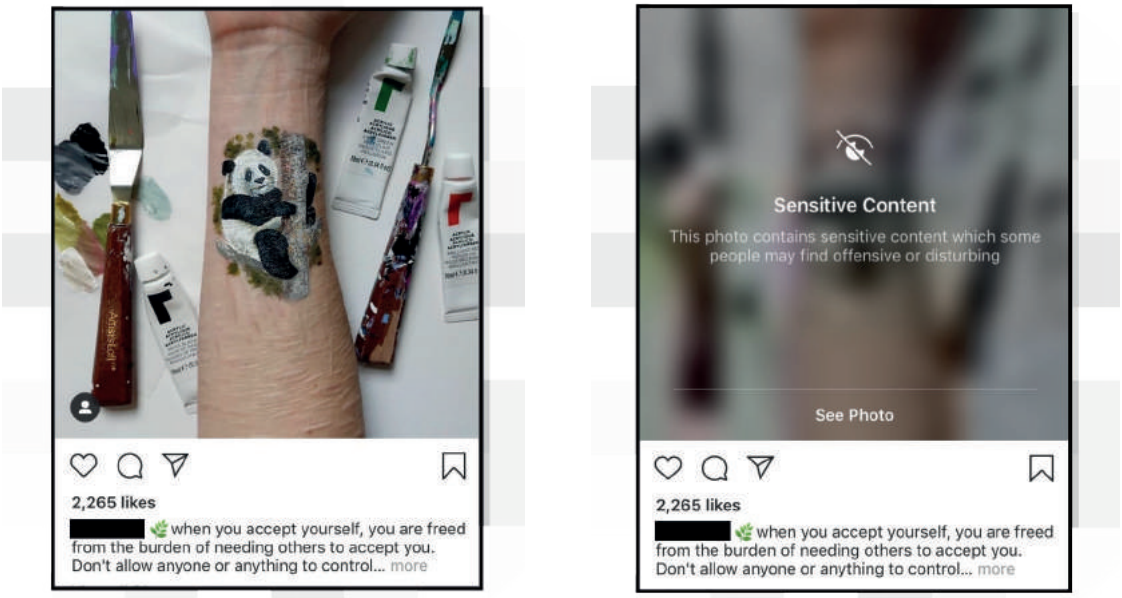
In practice, these rules have resulted in the removal many pieces of content from people

62 Community Guidelines, Instagram, https://www.facebook.com/help/instagram/477434105621119/
63 Ibid

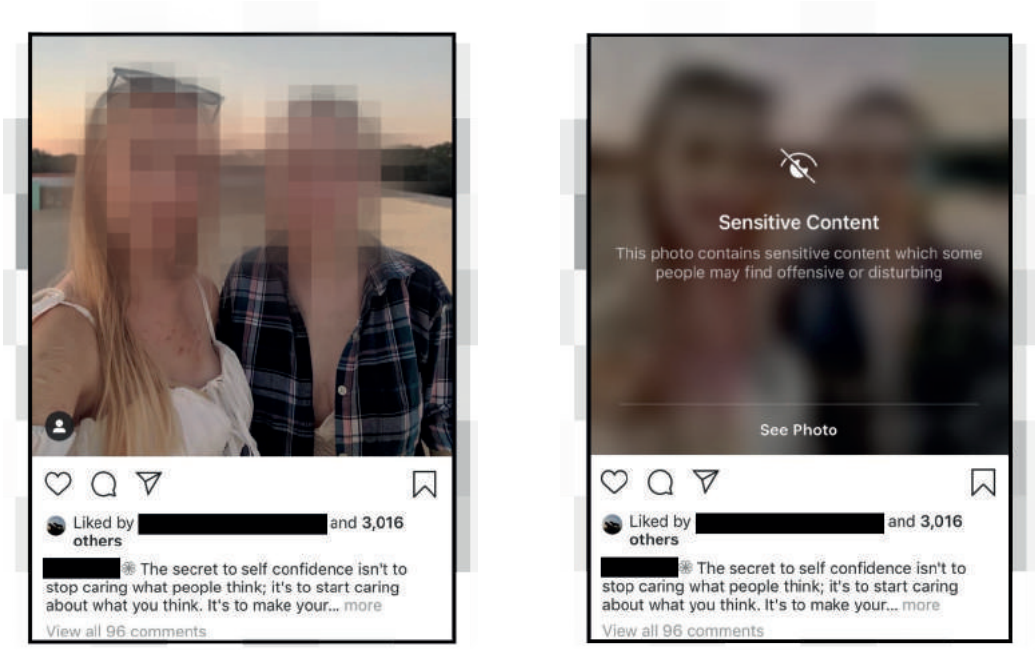
who have experienced self-harm and want to share their stories. It must be recognised that censorship and the social ostracisation it can cause may also have an impact on the mental health and wellbeing of individuals affected.



This picture, posted by an Instagram user, showing their scars healing and discussing recovery from self-harm, was censored by the platform.

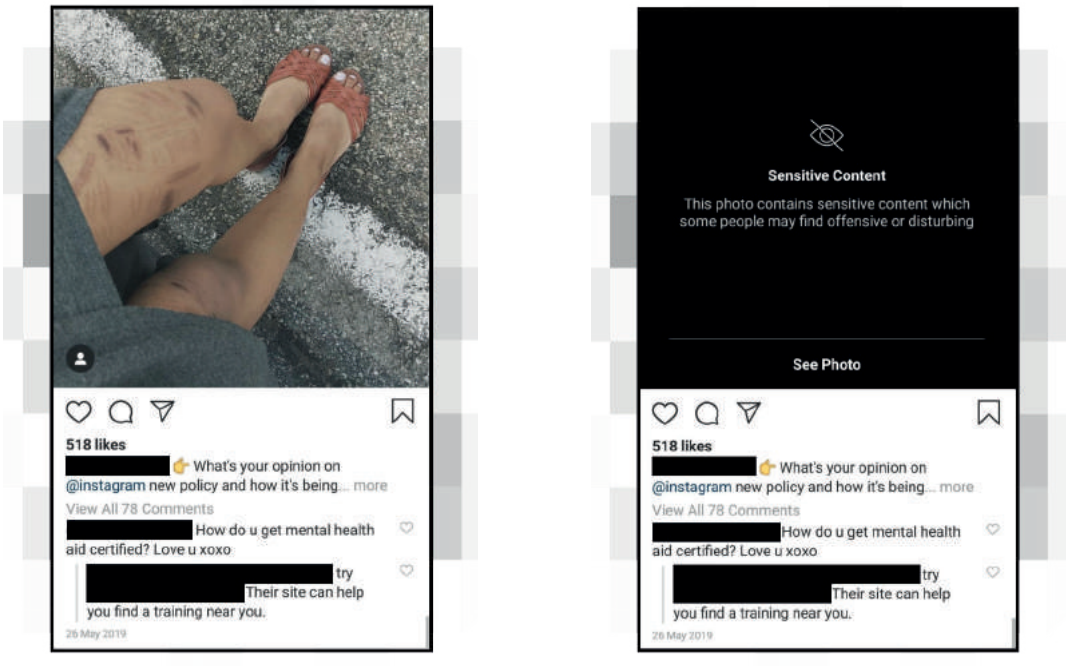


Instagram obscured this user's photo of body art due to the presence of scars.

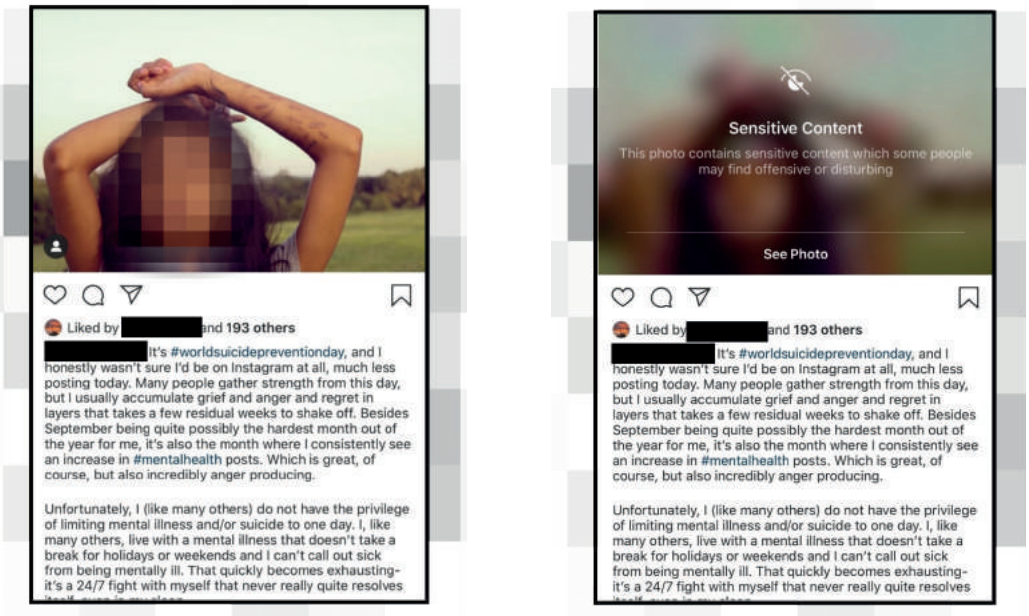


I

nstagram's algorithm, which censors content containing self-harm scars, labelled this picture of a young woman and her friend as "offensive or disturbing".



In this image, an Instagram user mentions the platforms algorithmic censorship of self-harm survivors' scars, only for the site to censor the image.

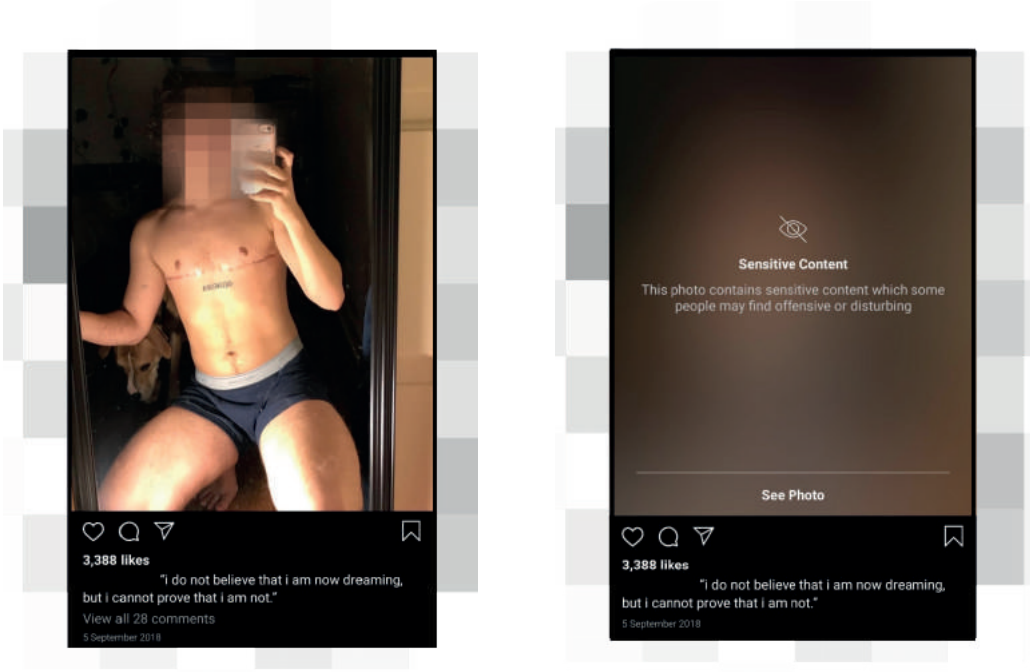


This young woman's photo, marking World Suicide Prevention Day, was censored because her arms are scarred.

Far from supporting those who have experienced serious mental illness, Instagram's censorship is profoundly stigmatising.

Instagram's censorship of people who have courageously shared their experiences of self-harm on the platform actively perpetuates the stigma around self-harm. Far from creating an inclusive, welcoming environment for people who have experienced trauma or mental ill health, the platform actively discriminates against them. Publicly labelling people's bodies as "offensive or disturbing" and obscuring them from view in their social circle is not only likely to cause severe distress but will also have a chilling effect on others who have had similar experiences and are made to feel unable to share their personal photos or stories online. Far from emboldening its users, Instagram's censorship in this case exacerbates old taboos around mental health.

Instagram's automated censorship of scars has also caused discriminatory treatment of trans people. We've identified many cases where photos of trans people have been censored as "sensitive" and potentially "offensive or disturbing" owing to surgery scars. This is unjustified, stigmatising and potentially very harmful for individuals affected.



This photo of a young trans man, showing scars, was censored by Instagram.

It Matters

It Matters is a global network of young people, working with fellow creatives and campaigners to explore mental health.

It Matters gave us the testimonial below regarding censorship, social media and mental health related content.

Since the tragic suicide of Molly Russell, Instagram and other Big Tech companies have quite rightly been scrutinised for failing to take down material that glamourises self-harm and suicide. Unfortunately, however, so much of the dialogue surrounding this issue has lacked nuance. Well-meaning calls to tackle cyber-bullying, pro-ana accounts and self-harm are being used to justify an authoritarian drive to silence young people and expand the powers of both Big Tech and the Government.

As the generation most affected by these issues, we find it alarming how our perspectives have routinely been excluded from this discussion. In doing so, the approach to tackling harmful content has more often been more fearfully reactive rather than sensibly proactive.

Instagram has been widely praised for taking down images related to self-harm. This knee-jerk response, however, has fuelled more stigma against mental illness and harmed the very young people it has sought to protect. Many young people are having harmless photos of themselves being deemed 'offensive' by Instagram's algorithm. As a consequence, these photos have been automatically blurred or removed. We have found that many of these photos have not promoted self-harm. Instead, they are photos of young people who have recovered from self-harm, continuing to live their ordinary lives. The scars in these photos are not fresh. In many instances, these scars were not visible at all.

Due to such an intense level of censorship, young people have reported feelings of being 'bullied', 'isolated' and 'humiliated' by Instagram's censorship. There is thus a real danger that Instagram's algorithms have triggered young users, causing them to relapse. We are particularly concerned about how such censorship is encouraging young people to seek more harmful areas of the internet.

The lack of transparency as to how these algorithms operate is worrying. Who, in Instagram, gets to decide what is 'harmful' material and what is not? According to our own research, we have found the following:

- ***Instagram is giving conflicting and misleading messages***

  When a post is censored, it is often accompanied with a message that 'someone thinks that you need help', implying that the post had been reported by a fellow user. However, we have found scenarios where a user has posted an identical image on two separate accounts. One of these accounts was a private account with no followers. Yet on both accounts, the image had been taken down, with the accompanying message that 'someone thinks that you need help'.

- ***Instagram's algorithms are blacklisting mental health advocates***

  Young mental health influencers, who have promoted recovery, have long suspected that Instagram has blacklisted them. This increasingly seems to be the case. Many users have reported to see the message 'you have chosen to see fewer posts like this' on young people's accounts, as well as mental health pages. Ironically, for many of these users, no such 'consent' was given.

- ***Instagram is censoring trans bodies***

  Young, trans people seem to have also been censored, due to their post-op scars. This again is legitimising and reinforcing ignorance.

According to the Government's proposed Online Safety Bill, Big Tech companies will be encouraged to enforce an even greater crackdown on material that they deem 'harmful'. Given Instagram's current lack of transparency, we fear that blanket censorship would become a convenient cop-out from instead investing more time and resources in developing more sensitive policy.

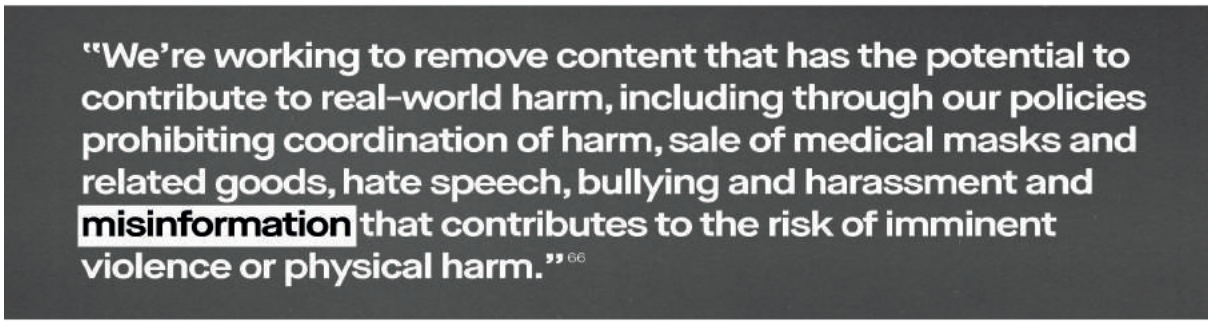
We know from our network of young influencers and young mental health advocates that there is plenty of data already out there in the public domain. However, this data is not being fully utilised, as there is considerable lack of engagement between Big Tech and its users.

We feel that a greater, more transparent collaboration between researchers, Big Tech companies and young users is the best way forward. That way, young people will feel more empowered and this vital conversation will not be periodically side-lined, as it often is.

Instagram and COVID-19

Instagram adopted parent company Facebook's community guidelines on COVID-19 content in January 2020 after the WHO "declared the coronavirus a public health emergency of international concern".[64] In practice, this means that content related to COVID which is deemed to constitute "misinformation" that could cause "imminent violence or physical harm"[65] is removed. However, this definition is applied very loosely.

Instagram's community guidelines page states:



This policy was updated and broadened substantially alongside Facebook on 8th February 2021, to include "additional debunked claims about the coronavirus and vaccines" including some claims about the origins of the virus.[67] Where content is considered to be "misinformation" but does not pose a risk of "harm" and does not fall within the specified list of "debunked claims", it may be assessed by independent fact-checkers and subsequently labelled. Like Facebook, fact-checking is based on guidance from the WHO and other health authorities.

However, labelling of this kind has not been limited to posts which are believed to contain "misinformation". Instagram's concern about misleading content on the site has led it to include a label, directing users to the websites of health authorities on all posts concerning COVID-19 or vaccines.

64 Facebook, Keeping People Safe and Informed About the Coronavirus, https://about.fb.com/news/2020/12/coronavirus/

65 Facebook Community Guidelines, Violence and incitement, Violence and criminal behaviour, https://www.facebook.com/communitystandards/credible_violence

66 Community Guidelines, Instagram, https://www.facebook.com/help/instagram/477434105621119/

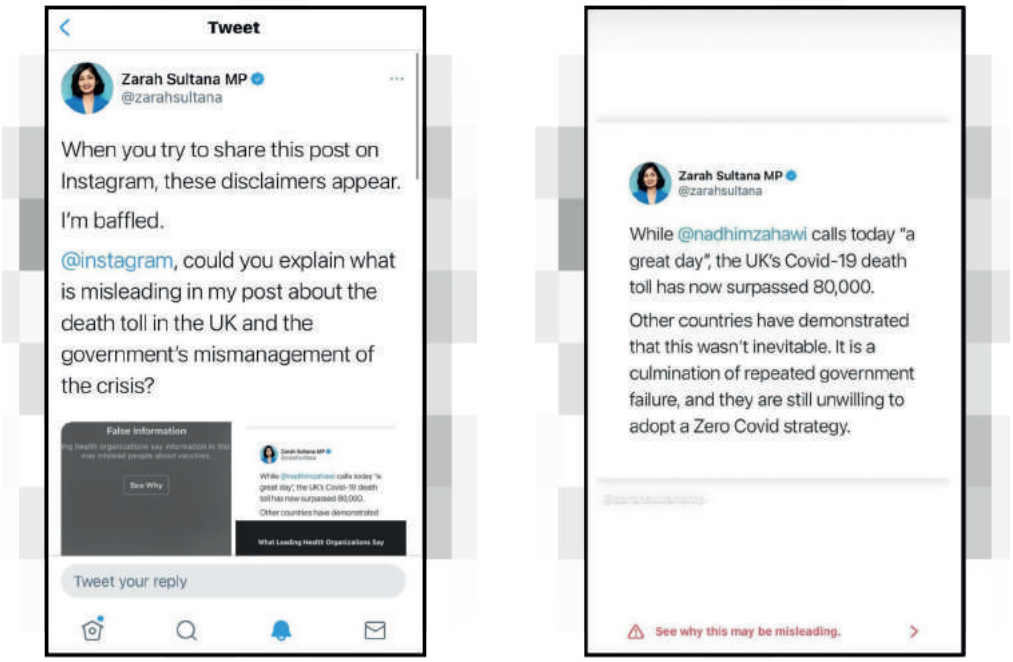
67 Facebook, An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19, https://about.fb.com/news/2020/04/covid-19-misinfo-update/#removing-more-false-claims

Information labels on Instagram, that are applied to all posts related to COVID-19 and vaccinations

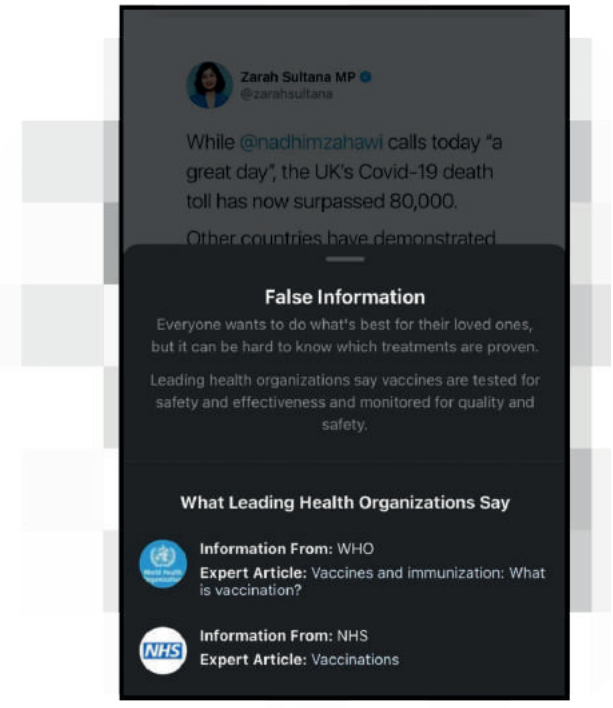
Instagram flagged a post by Labour Member of Parliament Zarah Sultana about the Government's management of the pandemic as "misleading". The MP for Coventry South shared a picture which stated:

> "While @nadhimzahawi calls today "a great day", the UK's Covid-19 death toll has now surpassed 80,000. Other countries have demonstrated that this wasn't inevitable. It is a culmination of repeated government failure, and they are still unwilling to adopt a Zero Covid strategy." [68]

68 Zarah Sultana, Twitter, 2021, https://twitter.com/zarahsultana/status/1348018234285584385



Zarah Sultana MP's Instagram post, criticising the Government's handling of the pandemic, was censored and marked as "misleading".

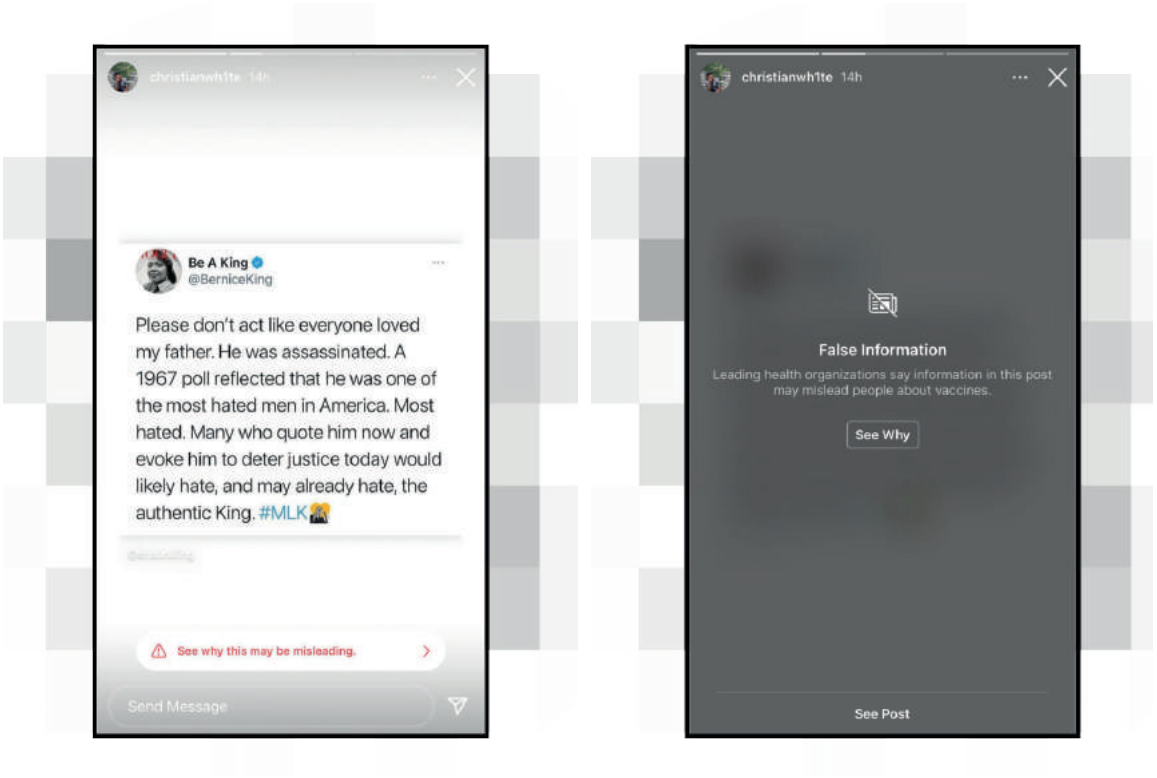


Zarah Sultana MP's Instagram post, criticising the Government's handling of the pandemic, was censored and marked as "misleading" despite not containing any factually incorrect information.

According to Instagram, the post contained "false information". The platform's content moderation system also stated that "Leading health organisations say information in this post may mislead people about vaccines." These are serious allegations to make about anyone's speech – particularly the communications of a Member of Parliament. However, the post objectively did not contain false information or any information regarding vaccines. This example demonstrates the platforms' increasing willingness to arbitrate the speech of democratically elected politicians, even on matters of public policy.

Once again, the policing of these terms of use through automated systems resulted in censorship, democratic interference, and even the complete mis-application of the rules themselves.

Instagram's automated Covid-19 policy has even resulted in political censorship of entirely unrelated speech. This was evident when a post by American minister Bernice King, daughter of civil rights leader Martin Luther King, marking MLK Day in 2021 was censored. Instagram flagged the viral post as "False information" that "may mislead people about vaccines", despite the post containing no information of this kind.



The extreme inaccuracy of this kind of automated content moderation demonstrates the problems with intervening in lawful communications at scale, and the real harms such arbitrary censorship can have – particularly when it is applied arbitrarily and incorrectly to posts regarding political leaders.

"The inevitable [redacted] consequence of the [redacted] online 'harms' agenda would be the censorship [redacted] of the most [redacted] marginalised groups in society. Expression [redacted] that is unpopular, [redacted] controversial or counter-cultural will be, if [redacted] not directly [redacted] targeted, seen as merely collateral damage under risk-averse, politically-influenced [redacted] content policies."

BIG BROTHER WATCH

Twitter

Twitter and hate speech

Twitter's "Hateful Conduct Policy" informs users they "may not promote violence against or directly attack or threaten other people" based on a series of characteristics.[69] It should be noted that promoting violence against or directly threatening others will ordinarily pass a criminal threshold in the UK regardless of whether it is directed to people with protected characteristics or not. However, further reading of Twitter's hateful conduct policy reveals it extends far beyond the promotion of violence, attacks and threats.

Twitter's "protected characteristics", like other platforms, exceed those defined in UK law. Mirroring protected characteristics included in the UK's hate crime definition, they include disability, race (and national origin and ethnicity), religion and sexual orientation; but deviating from domestic law, they include two categories of 'gender identity' and 'gender' as opposed the protection of 'sex' and 'gender reassignment' which is protected by statute.[70] Additionally, Twitter includes age; like Facebook it includes serious disease; and like both Facebook and YouTube it additionally includes caste as a category protected from 'hateful conduct'.[71]

Twitter says it reviews reportedly "hateful content" to establish "whether the intention is to abuse an individual on the basis of their protected status, or if it is part of a consensual conversation".[72] However, the examples of Twitter's "hateful conduct" enforcement that we have reviewed show that content that is neither abusive nor targeted at another individual is frequently censored, particularly in relation to sex and gender.

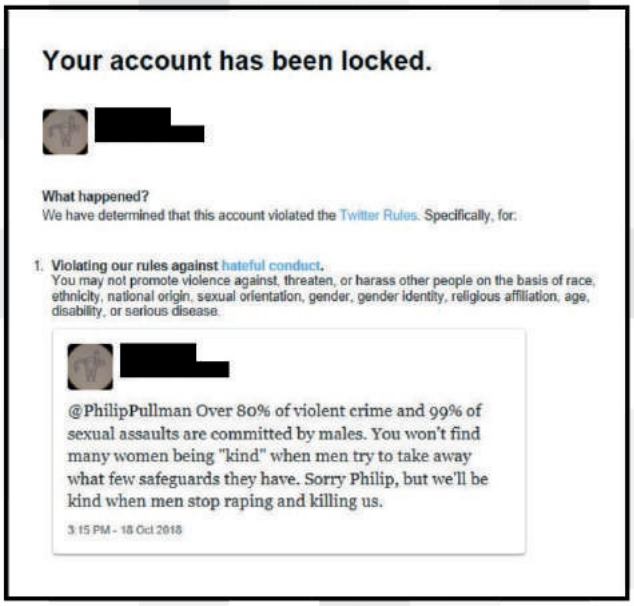
Elaborating on what types of speech are included within the 'Hateful Conduct Policy', Twitter prohibits a number of categories including to 'incite fear... about a protected category'.[73] The policy is written to prohibit content that has malintent, banning 'content intended to incite fear or spread fearful stereotypes', but deducing intent from short-

69 Twitter, Hateful Conduct Policy https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
70 Ibid.
71 Ibid.
72 Ibid.
73 Ibid.

form content like tweets can sometimes be very difficult and rely on interpretation.[74] The policy is written to include 'asserting that members of a protected category are more likely to take part in (...) illegal activities'.[75] Given that 'gender' is included as a protected category, this means that stating facts about male crime rates has been frequently considered hateful conduct that is in breach of the policy. We have seen an extraordinary number of examples of women being censored, suspended or banned from Twitter for stating such facts.

Men and crime

Since gender is considered a protected characteristic on Twitter, 'generalisations' about men, especially those that may incite fear or fearful stereotypes, are banned. Discussion and fact-sharing of male crime rates among women have fallen squarely in Twitter's policy against 'asserting that members of a protected category are more likely to take part in (...) illegal activities', and resulted in many women's accounts being temporarily limited, suspended or banned.

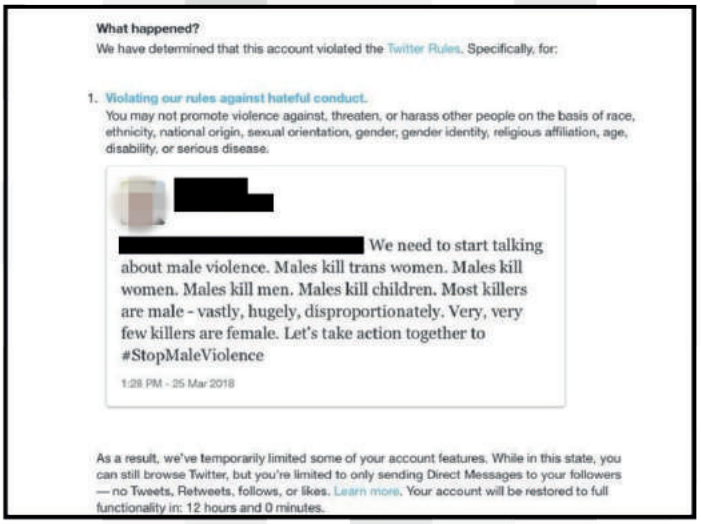


74 Twitter, Hateful Conduct Policy https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

75 Ibid.

Woman locked out of Twitter for citing male crime statistics

In this example, a woman's account was locked for making assertions which align with ONS data[76] about male crime to author Philip Pullman, as this violated Twitter's rules against "hateful conduct" towards men.



Woman censored and restricted on Twitter for post about male violence

In this example, a woman's tweet was removed and her account limited after her explanation that the majority of violent crime is perpetrated by men was deemed by Twitter to violate the hateful conduct policy.

76 Office for National Statistics, The nature of violent crime in England and Wales: year ending March 2018, 2019, https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/articles/thenatureofviolentcrimeinenglandandwales/yearendingmarch2018#what-do-we-know-about-perpetrators-of-violent-crimes and Sexual offences in England and Wales: year ending March 2017, 2018, https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/articles/sexualoffencesinenglandandwales/yearendingmarch2017

Woman restricted on Twitter for 7 days after posting legal definition of rape

In this example, a woman was stopped from tweeting for 7 days after posting about the definition of rape in UK law. This breached the "hateful conduct" policy.

Woman suspended on Twitter for citing legal definition of rape

In this example, a woman responded to a user who asked her about women and rape by saying that UK law states that women cannot rape. Her account was suspended and the tweet deleted. After appeals, including by our Director to Twitter's policy team, the suspension was lifted and the tweet reinstated. These cases show examples of where Twitter opt for censorship over a free discussion around subject areas which are difficult or uncomfortable.

Misgendering

Another prohibited category within the hateful conduct policy is 'repeated and/or non-consensual slurs, epithets, racist and sexist tropes, or other content that degrades someone'.[77] Whilst racist tropes are clear to identify, the more general categories of 'non-consensual slurs' and 'content that degrades someone' are clearly vague and will rely to a great degree on the subjective judgment of moderators.

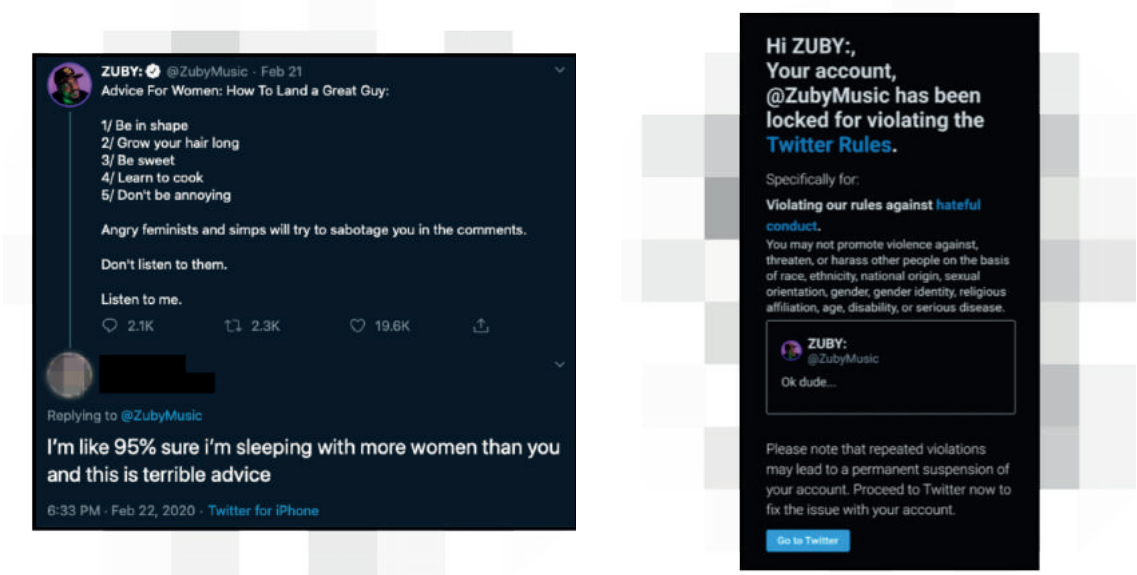
Within this same category of degrading content is a rule against 'misgendering'. To 'misgender' someone, especially a transgender person, is to refer to them using a word, especially a pronoun or form of address, that does not reflect the gender with which they identify.[78] The policy also prohibits 'deadnaming of transgender individuals' - that is, referring to transgender people by their birth name after they have specified their chosen name.[79]

Misgendering and deadnaming touch on highly sensitive topics, and both can be very hurtful to trans people. However, misgendering and deadnaming are not criminal offences in the UK. Indeed, compelling speech (e.g. the use of certain pronouns) could raise tensions around freedom of expression. It is possible that misgendering and deadnaming could be involved in crimes against transgender individuals, but a criminal offence (such as harassment or a communications offence) would have to be involved.

However, we have seen many examples of this well-intended policy having negative consequences including undue censorship of women and transgender people.

77 Twitter, Hateful Conduct Policy https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
78 Lexico, Definition – 'misgender', https://www.lexico.com/definition/misgender
79 Twitter, Hateful Conduct Policy https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

Twitter's hate speech and misgendering policy has been enforced where users' use of common forms of address are interpreted as entailing misgendering. This typically involves the words 'guy' and 'dude'.



A man was suspended by Twitter for saying "dude" to a trans woman

In the case above, the rapper and podcast host Zuby responded "ok dude" to another user, who is a trans woman. She reported the tweet for "misgendering" and Zuby was subsequently suspended by Twitter.

Zuby told the Washington Examiner:

> *"It wasn't even a gendered statement, not that it should even matter. I used it as a synonym for like, 'Yeah, whatever.' I didn't go and research the person I was responding too. Perhaps I should have because they're an antifa activist who spends their time trying to doxx people and get people kicked off platforms that they don't like."*[80]

Zuby appealed the suspension, but was unsuccessful and could only regain access to his account by deleting the tweet:

> *"I could either delete the tweet and my account would be reinstated after 12 hours, or if I thought they made a mistake, I could appeal, so I decided to appeal.*

80 'Ok dude': Twitter suspends rapper Zuby for 'hateful' tweet at transgender antifa activist – Spencer Neale, Washington Examiner, 27 February 2020: https://www.washingtonexaminer.com/news/ok-dude-twitter-suspends-rapper-zuby-for-hateful-tweet-at-transgender-antifa-activist

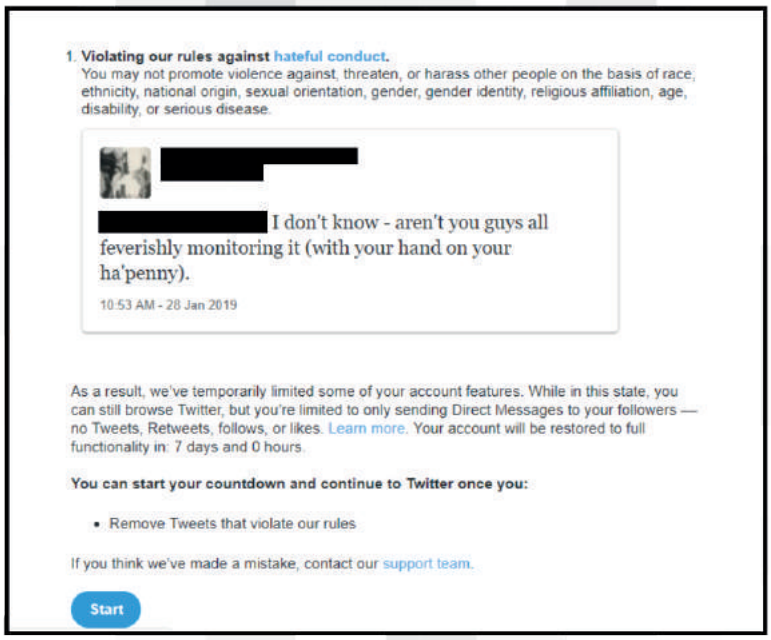
This morning, I found that the appeal had been rejected and the only course of action from then was to delete the tweet in question. There was no other option."[81]

Whilst being referred to as "dude" may have been upsetting to the individual, it does not follow that censorship is a proportionate response. It is not even clear that the word "dude", which is often used in a non-gendered way, breached the misgendering policy in this case, unless the policy vests power primarily in the subjective interpretation of the reporter. This case demonstrates how such granular and politicised censorship can be counter-productive and inflame debates, leading to greater controversy.

Woman locked out of Twitter for referring to someone as "dude"

In this case, a feminist writer was locked out of her account for using the word "dude". She told us she had dealt with the "double standard" of having her tweets deleted whilst Twitter did not deal with serious online harassment against her (resulting in her taking legal action about said harassment). As a consequence, she deleted her Twitter account.

81 'Ok dude': Twitter suspends rapper Zuby for 'hateful' tweet at transgender antifa activist – Spencer Neale, Washington Examiner, 27 February 2020: https://www.washingtonexaminer.com/news/ok-dude-twitter-suspends-rapper-zuby-for-hateful-tweet-at-transgender-antifa-activist



Woman censored and suspended on Twitter for using the word "guys" and "ha'penny"

In this case, a woman uses the terms "guys" and "ha'penny" (slang for 'privates', typically female) to a Twitter user who does not use a full name but states their pronouns as "she/her". Twitter found the tweet to fall under "hateful conduct" and she could only regain access to her account if she deleted the tweet and then took a seven-day suspension. Her account has since been deleted.

'Cis' and 'TERF'

Terms such as 'cis' and 'TERF' have also featured in Twitter's misgendering and hateful conduct policy.

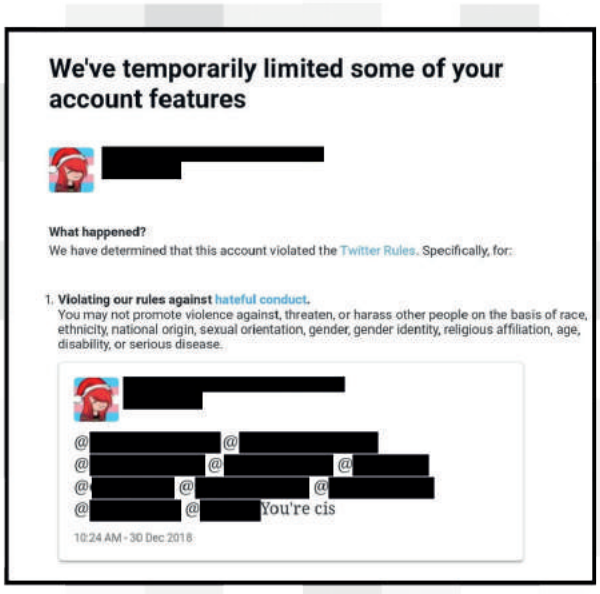
'Cis' is a Latin prefix meaning 'on this side of'. It is used in some discourse on gender to refer to people who are not transgender. The word cisgender is defined as "Denoting or relating to a person whose sense of personal identity and gender corresponds with their birth sex."[82] Some people view that the word 'cisgender' implies that people who are not transgender positively identify with the gender role associated with their sex, and therefore they oppose it. It has also been argued that, the words 'cis' and 'cisgender'

82 Lexico, Definition – 'cisgender' https://www.lexico.com/en/definition/cisgender

assume a gender identity that may not be held, and even objected to, by the individual it refers to. As a result, use of the word 'cis' started to appear in Twitter misgendering enforcement.

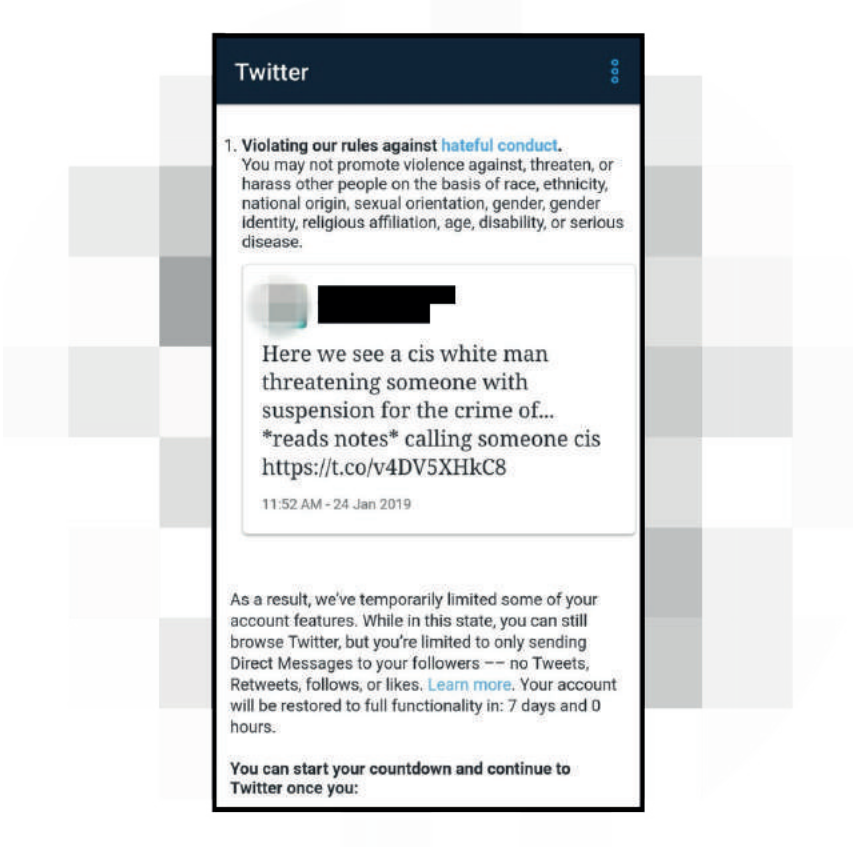
Whether one personally uses the term 'cis' or not, it is part of a political lexicon and is central to a taxonomy of gender in some discourses on gender identities. Twitter's policing of the term stifles such legitimate discourse.

The acronym 'TERF' can also be found in gender debates, although it is informal and typically used as a slur or term of abuse. 'TERF' stands for 'trans-exclusionary radical feminists' and is often used to refer to people, especially women, who are either transphobic or who advocate for single-sex spaces. Because the term is often used in the context of abuse, Twitter took enforcement action against tweets with the word 'TERF", likely on the basis that it falls within the hateful conduct policy against 'non-consensual slurs'. Regardless of whether people object to the term TERF, Twitter's attempted enforcement against the term constituted interference in a political lexicon that was unjustified and censorious.



Journalist's account restricted on Twitter for after saying "You're cis"

In this case, a trans journalist's account was temporarily limited after she said "You're cis".

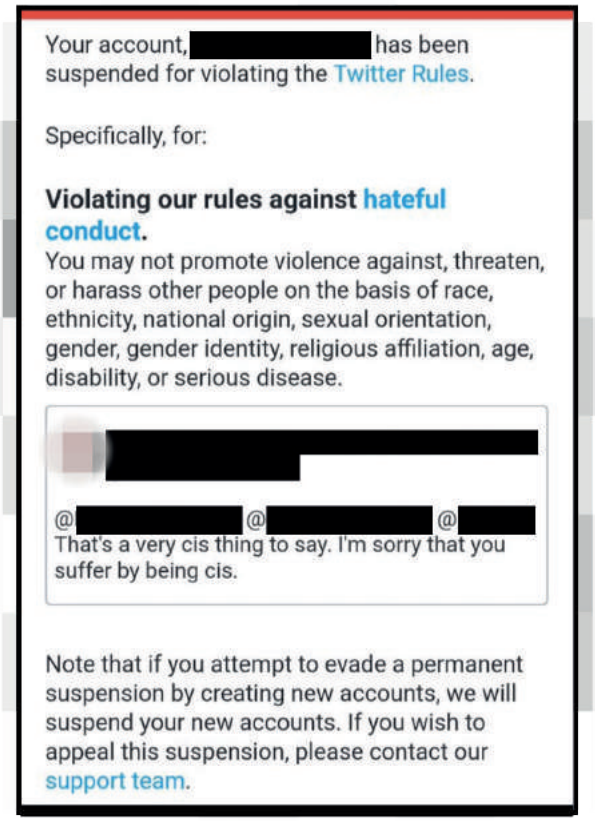


User's account restricted for calling another user "cis"

In this case, a user's account was temporarily limited for calling a male user 'cis'. The male user had threatened to report other users who referred to him as 'cis'. The enforcement action required the user above to delete the tweet and then begin a 7-day suspension before regaining full account access. The account no longer exists.

Woman's account restricted on Twitter for saying she was less interested in talking to "cis men"

Similarly, this woman's account was limited after she tweeted that she was "less interested in talking to cis men about women's issues". The man in the thread said that he reported the tweet as he doesn't identify as 'cis'.



Twitter user suspended for use of the word "cis"

This account was suspended are posting a tweet on a thread saying "That's a very cis thing to say" and intimating that another user was 'cis'.

Twitter has no way of knowing the gender of unverified account holders – nor would it be appropriate for Twitter to adjudicate on individuals' gender. However, this means that the misgendering policy is at individual users' behest and, if abused, could be exploited to censor and silence others.

Twitter and COVID-19

In May 2020, Twitter introduced a new policy regarding COVID-19 related content on its platform.

Twitter's policy states:

> *"Content that is demonstrably false or misleading and may lead to significant risk of harm (such as increased exposure to the virus, or adverse effects on public health systems) may not be shared on Twitter."*[83]

It goes on:

> *"In addition, we may label Tweets which share misleading information about COVID-19 to reduce their spread and provide additional context."*[84]

"Harm" is not defined, although Twitter's rules go further and specify five areas of prohibited content. These are:

- False or misleading information about the nature of the virus.
- False or misleading information about the efficacy and/or safety of preventative measures, treatments, or other precautions to mitigate or treat the disease.
- False or misleading information about official regulations, restrictions, or exemptions pertaining to health advisories.
- False or misleading information about the prevalence of the virus, or risk of infection or death.
- False or misleading affiliation.[85]

When considering rules for online platforms which govern "misleading content", it is important to consider distinctions between what different communications may entail. According to the Cambridge English Dictionary, disinformation is "false information spread in order to deceive people."[86] It is important to note that legislation already exists

83 Twitter, COVID-19 misleading information policy, https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy
84 Ibid.
85 Ibid.
86 Cambridge Dictionary, Definition – disinformation, https://dictionary.cambridge.orgdictionary/

in the UK to determine how the wilful spread of false information can be dealt with. The Communications Act 2003 prohibits the sending of a communication which is known to be false which causes "annoyance, inconvenience or needless anxiety to another".[87]

"Misinformation" on the contrary, as defined by the Cambridge English Dictionary, is "wrong information, or the fact that people are misinformed".[88] Arguably, a strict prohibition on "misinformation", whether through corporate terms and conditions or by law, is nigh-on impossible nor desirable.

The strict nature of Twitter's rules makes the company's enforcement more akin to truth arbitration than the protection of public safety, penalising those whose views or understanding do not align with those of authorities. In any free and fair democracy, citizens should have the right to speak freely and no supreme censor should hold power to adjudge the validity of their speech unless it is illegal and/or falls outside of the bounds of speech protected by human rights frameworks. Movement towards technocratic speech arbitration online where all statements must be empirically verified in order for them to be permissible is dangerous and not in the spirit of a democracy.

Moreover, Twitter states that "false or misleading" information about areas of public health policy could be either labelled or removed. This includes content around:

- Personal protective equipment (PPE) such as claims about the efficacy and safety of face masks to reduce viral spread;
- Local or national advisories or mandates pertaining to curfews, lockdowns, travel restrictions, quarantine protocols, inoculations, including exemptions from such advisories or mandates;
- The capacity of the public health system to cope with the crisis.[89]

Former advisor to President Trump, Scott Atlas fell afoul of these rules when he disputed the efficacy of masks. Atlas claimed that masks do not work, citing a study by the Oxford academic Professor Carl Heneghan. The Tweet was removed by the platform.

---

english/disinformation

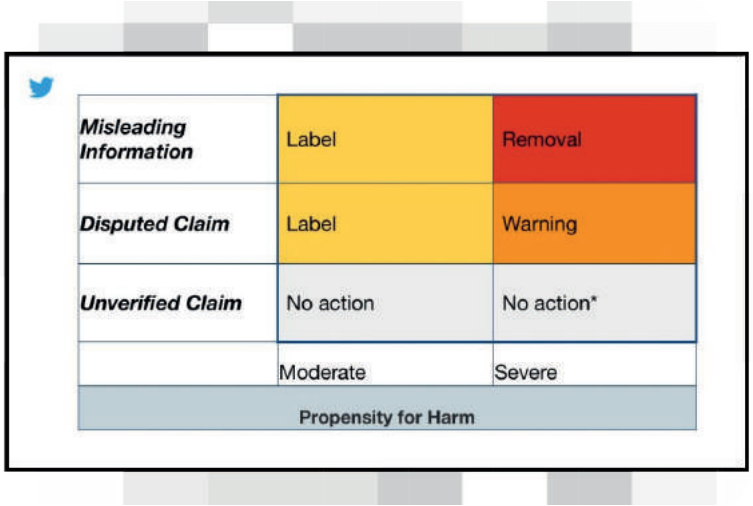
87 Communications Act 2003, S. 127, https://www.legislation.gov.uk/ukpga/2003/21/section/127

88 Cambridge Dictionary, Definition – misinformation, https://dictionary.cambridge.org/dictionary/english/misinformation

89 Twitter, COVID-19 misleading information policy, https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy

The former President's advisor's tweet was removed for disputing the efficacy of masks.

Twitter states that in circumstances where content removal is not deemed necessary the platform "may provide additional context on Tweets sharing the content where they appear on Twitter".[90] Content which does risk "harm" will be removed and the users' account may be temporarily suspended. According to Twitter, multiple violations of these rules could result in permanent account suspension.[91]



Twitter's matrix for enforcement of "misinformation"

90 Twitter, COVID-19 misleading information policy, https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy

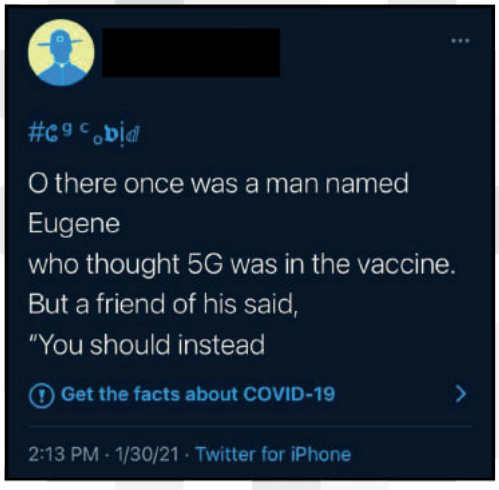
91 Ibid.

'Not a violation of this policy'

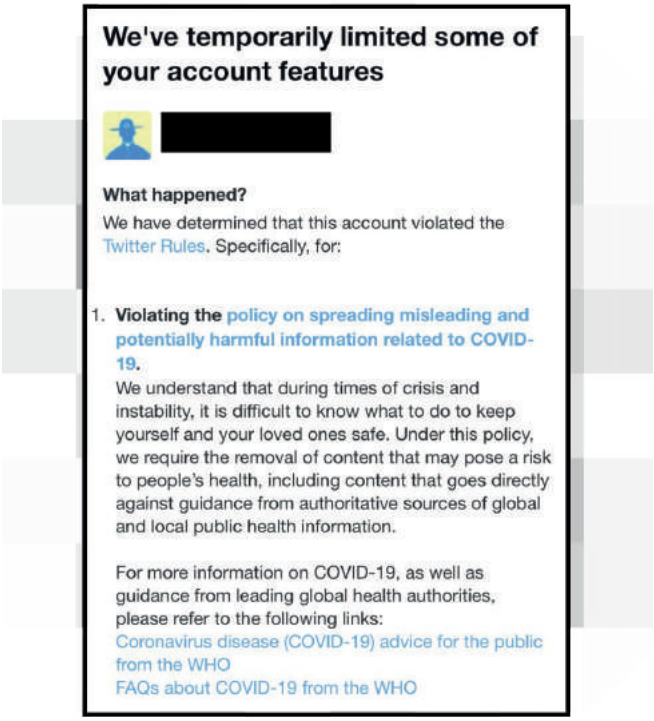
Twitter provides a number of content categorisations which it says will not violate this policy. Amongst those areas listed is:

- Strong commentary, opinions, and/or satire, provided these do not contain false or misleading assertions of fact.[92]

However, this has proved not to be the case in practice. Upon noticing that Twitter was labelling all tweets which contained a hashtag relating to COVID-19, one Twitter user decided to write a satirical Limerick, containing the hashtag, which concluded with the automated Twitter label itself. The account was temporarily suspended.



92 Twitter, COVID-19 misleading information policy, https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy



This Twitter user's account was suspended for posting a satirical Limerick including a hashtag about COVID-19 that invoked automated enforcement.

Another area Twitter says is not covered by this policy is "counterspeech". The platform states:

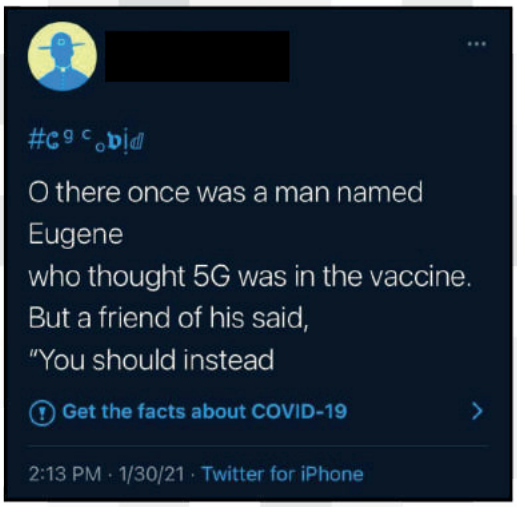
> *We allow for direct responses to misleading information which seek to undermine its impact by correcting the record, amplifying credible information, and educating the wider community about the prevalence and dynamics of misleading information.*[93]

While more speech is a healthy antidote to combatting lies and falsehoods, this only works in practice if the initial claim which is perceived to be false is not censored or deleted.

Finally, while not yet acknowledged by Twitter, credible reports have emerged of the platform pre-emptively preventing the posting of content where it is deemed "harmful". This is a worrying development in the field of content moderation. Rather than moving to

93 Twitter, COVID-19 misleading information policy, https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy

tackle illegal or dangerous content on the site retrospectively, the pre-screening content before it is even posted further narrows the permissibility of speech and contravenes generally accepted prohibitions on platforms' general monitoring of content.



This tweet regarding vaccines and sharing a link to a Guardian news article was pre-emptively blocked from being posted.

Twitter and political censorship

The suspension of former President Donald Trump's Twitter account will go down in history as a moment that changed the relationship between politics and technology. The rights and wrongs of Twitter's actions at various points leading up to the former President's suspension can be disputed, but it is without question that the case raises profound questions about the accountability and decision making of tech giants like Twitter, who now have the power to suppress world leaders.

In response to scenes of violence at the Capitol building in Washington, Twitter took the decision to remove three Tweets issued by Trump for "repeated and severe violations of our Civic Integrity policy."[94] amid fear that his rhetoric was further inciting unrest.

Twitter stated that "Future violations of the Twitter Rules, including our Civic Integrity or Violent Threats policies, will result in permanent suspension of the @realDonaldTrump account."[95]

In the days that followed the events at the Capitol Building, Trump issued two further tweets which resulted in his permanent suspension from the platform.

On 8th January, 2021, the then President tweeted:

> *"The 75,000,000 great American Patriots who voted for me, AMERICA FIRST, and MAKE AMERICA GREAT AGAIN, will have a GIANT VOICE long into the future. They will not be disrespected or treated unfairly in any way, shape or form!!!"*[96]

94 Tweet from Twitter Safety, 2021, https://twitter.com/TwitterSafety/status/1346970430062485505

95 Tweet from Twitter Safety, 2021, https://twitter.com/TwitterSafety/status/1346970432017031178?s=20

96 Twitter, Permanent suspension of @realDonaldTrump, 2021 https://blog.twitter.com/en_us/topics/company/2020/suspension.html

He tweeted again:

> *"To all of those who have asked, I will not be going to the Inauguration on January 20th."*[97]

Twitter's assessment was that in the context of events, the two tweets were a "violation of the Glorification of Violence Policy"[98] on the platform. However, the assessment relied on the company's view of how others interpreted Trump's tweets rather than the content itself. According to Twitter's statement:

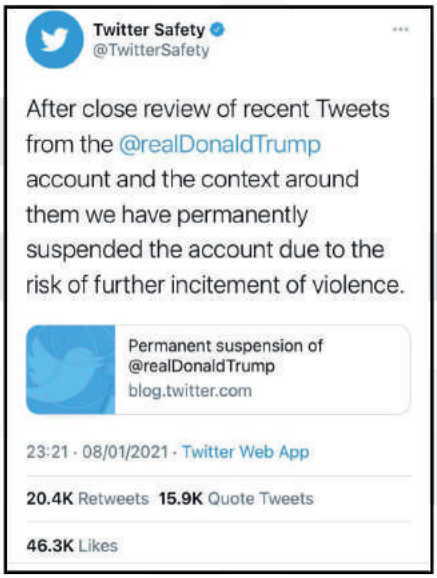
> *"The use of the words "American Patriots" to describe some of his supporters is also being interpreted as support for those committing violent acts at the US Capitol."*[99]

Additionally:

> *"The mention of his supporters having a "GIANT VOICE long into the future" and that "They will not be disrespected or treated unfairly in any way, shape or form!!!" is being interpreted as further indication that President Trump does not plan to facilitate an "orderly transition" and instead that he plans to continue to support, empower, and shield those who believe he won the election."*[100]

The remarkable action by Twitter signalled a radical reinterpretation of speech standards. Twitter's adjudication, leading to their censorship of the holder of the highest democratically elected office arguably in the world, relied upon an estimation of how some people may interpret the tweet, rather than what the President had actually said. This sets a precedent for new standards that can be applied to ordinary people on Twitter including those challenging power, not only those who hold it.

97 Twitter, Permanent suspension of @realDonaldTrump, 2021 https://blog.twitter.com/en_us/topics/company/2020/suspension.html
98 Ibid.
99 Twitter, Permanent suspension of @realDonaldTrump, 2021 https://blog.twitter.com/en_us/topics/company/2020/suspension.html
100 Ibid.



Then-President Donald Trump's Twitter account was infamously suspended, permanently, on 8th January 2021.

Twitter's action raises serious questions about whether content moderation on the platform is based on a clear set of rules or whether it is simply left to company executives to interpret speech and apply rules subjectively.

These incidents have resulted in heightened levels of intervention from Twitter content moderators. In the wake of the former President's disputation of the 2020 election result, the platform introduced a new "Civic integrity" policy which prohibits the sharing of misleading information regarding a "civic event".[101] While well intentioned, many of the rules specified within this policy, such as the posting of "misleading claims about long lines, equipment problems, or other disruptions at voting locations during election periods"[102] would not violate any UK law.

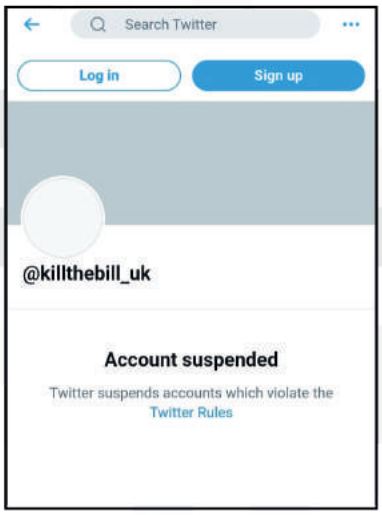
According to the platform, violations of these policies may simply result in tweet "labelling"; however, for "high-severity violations", the tweet may be removed and the user's account temporarily suspended.[103]

101 Twitter, Civic Integrity Policy, https://help.twitter.com/en/rules-and-policies/election-integrity-policy
102 Ibid.
103 Ibid.

It has been reported that Twitter suspended 70,000 accounts linked to the conspiracy theory group QAnon in the wake of the security breach at the Capitol.[104]

The platform also cracked down on left-wing political groups, suspending the accounts of a number of high-profile Antifa activists. According to a statement given to The Sun newspaper, the activists in question were suspended for violating the site's rules on "Platform manipulation and spam".[105] The removal of a large number of high-profile activists all from the same political movement, without reasonable qualification is deeply concerning both with regard to freedom of expression and political participation.

This kind of arbitrary political censorship was also seen in Britain where Twitter suspended the account of grassroots campaign group "Kill the Bill", who oppose the Government's Police, Crime, Sentencing and Courts Bill. The campaign group were given no explanation for the suspension of their account and the group set up another account which was also shut down.

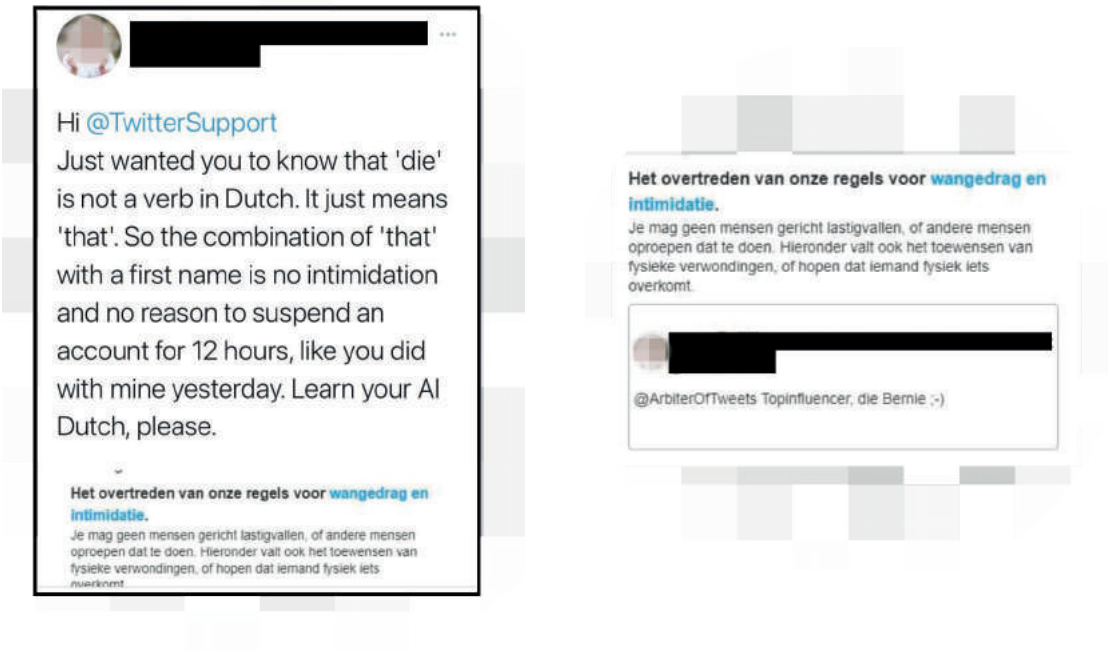


After a long suspension and the raising of this case by Big Brother Watch, the account was finally reinstated. However, in the interim, the Kill the Bill campaign were forced to set up a third account in order to promote their activities on the site. The removal of a campaign group from a major platform like Twitter, when they are campaigning against a live piece of legislation, is deeply damaging to the democratic process.

104 BBC News, Twitter suspends 70,000 accounts linked to QAnon, 2021 https://www.bbc.co.uk/news/technology-55638558

105 Knox, P. ANARCHISTS BLOCKED: Twitter suspends several popular Antifa accounts after Inauguration Day riots, The Sun, 2021, https://www.thesun.co.uk/news/13825038/twitter-suspends-antifa-accounts-inauguration-day-riots/

Amidst concerns about violence towards politicians, one Twitter user, tweeting about American politician Bernie Sanders in Dutch, saw her account suspended after content moderators read the words "die Bernie" (meaning "that Bernie" in English) and deemed the Tweet a violation of Twitter's "Abusive behaviour" rules. Whether a purely automated decision or human moderation, this demonstrates the very basic failures running through Twitter's enforcement processes.



Content moderators failed to realise that the tweet was written in Dutch and removed the post and suspended the user for a violation of Twitter's abusive behaviour policies.

YouTube

YouTube prohibits hate speech according to its own definition, which centres on the promotion of violence or hatred.[106] This includes dehumanising certain individuals or groups (e.g. with references to animals or insects), using stereotypes that promote hatred or treating such stereotypes as factual, or claims that certain individuals or groups are inferior. However, it is an expansive policy that also includes denial "that a well-documented, violent event took place", and "attacks" on a person's "emotional" attraction to another person. [107]

The prohibition on denying that a well-documented, violent event took place[108] appears to be designed to combat various conspiracy theories including Holocaust denial. Under UK law, there is no specific prohibition on Holocaust denial or so-called 'revisionism', although it can be prosecuted if it passes the threshold of "grossly offensive" communications (see R v Chabloz, 2019).

As with other platforms, the groups protected under YouTube's policy exceed those protected under UK hate crime law. They include the protected groups of disability, race (and nationality), religion, and sexual orientation but further protect the broad, perhaps limitless field of "gender identity and expression" rather than the characteristic of gender reassignment protected under UK law.[109]

Like Facebook, YouTube additionally protects sex, gender, caste and immigration status[110]. This means that content relating to 'exclusion' based on 'immigration status' is technically in breach of YouTube's policy, which arguably could put the immigration laws of every nation state in the world in breach of the policy. Furthermore, the platform's policy covers age, ethnicity, "victims of a major violent event and their kin", and veteran status.[111]

The creation of a protected characteristic for victims of major violent events appears reactionary, in response to conspiracy theories perpetrated on YouTube about school shootings in the US – which are typically dealt with as defamation or harassment under the

106 YouTube, Hate speech policy, https://support.google.com/youtube/answer/2801939?hl=en
107 Ibid.
108 Ibid.
109 Ibid.
110 Ibid.
111 Ibid.

law. This arguably confuses the public's understanding of what protected characteristics are and why they exist.

Furthermore, the creation of a protected status for veterans is unusual and arguably political. The policy does not state whether it is only US veterans who are protected or ex-armed forces personnel globally, but neither option qualifying as a protected characteristic can be clearly grounded in any equality rights analysis.

Importantly, YouTube reserves discretion to allow educational material that may fall under the hate speech policy, stating "we may allow content that includes hate speech if the primary purpose is educational, documentary, scientific, or artistic in nature".[112]

However, this exemption is not always applied where it should be. A number of journalists, educators and activists who document or challenge hate speech have had videos removed under the hate speech policy. This is partly a problem of policy, and partly enforcement – particularly automated enforcement which cannot accurately distinguish between content challenging or documenting hate speech and actual hate speech.

YouTube, hate speech and journalism

YouTube's introduction of a new hate speech policy in June 2019 led to a wave of censorship. The platform's attempt to eradicate right wing hate speech led to journalists who document the rise of neo-Nazism having their content removed.

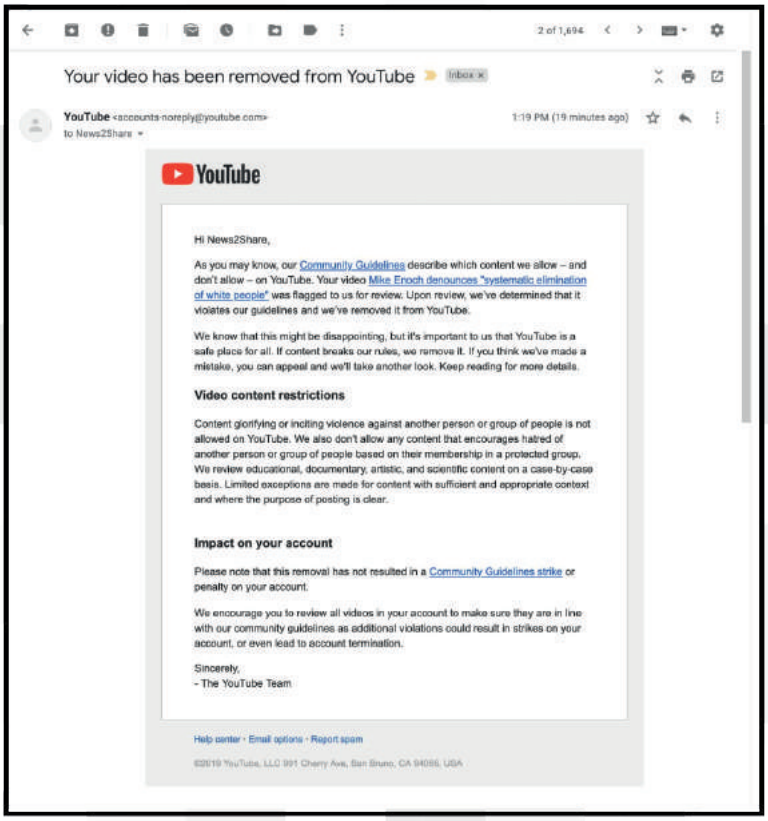
Ford Fischer is the Editor of News2Share, an online and largely YouTube based media outlet. His footage has been used for major broadcast documentaries, including PBS' "Documenting Hate" series. Fischer has been subjected to censorship and demonetisation in relation to several videos since the new hate speech policy was introduced.

His video of a high-profile white supremacist, Michael Peinovich (aka "Mike Enoch"), was removed by YouTube in June 2019. Fischer received an email informing him that content "glorifying or inciting violence" or that "encourages hatred" is not allowed on YouTube. Fischer complained that "while unpleasant, this documentation is essential research for history".[113] Indeed, this exact footage of his had already been used in the

112 YouTube, Hate speech policy, https://support.google.com/youtube/answer/2801939?hl=en
113 Ford Fischer, Twitter, 2019, https://twitter.com/FordFischer/status/1136334785297702912

PBS documentary series and his footage from the same event was used in the New York Times documentary "How an Alt-Right Leader Lied to Climb the Ranks."

Fischer commented, "YouTube and big tech censorship of journalists provides cover to the people they claim they're opposing."[114] However, despite his appeals, Fischer's video remained censored.



Ford Fischer's YouTube video of a white nationalist – which was used for broadcast documentaries – was removed from the platform for violating the hate speech policy.

These rules have not only impacted upon users in the UK. Earlier this year YouTube deleted the account of La Marea, a left-wing Spanish publication which had been reporting on anti-migrant vigilante groups. According to the platform, the reports had fallen foul of YouTube's hate speech guidelines and had "incited hatred".

114 Ford Fischer, Twitter, 2019, https://twitter.com/FordFischer/status/1136334788304998405

In both cases, the rules failed to distinguish between genuinely hateful or illegal content and the lawful and often important documentation of difficult and challenging issues.

The La Marea YouTube Channel was suspended for reporting on anti-migrant vigilantes.

YouTube and COVID-19

Like other platforms, YouTube introduced new policies in the wake of the COVID-19 pandemic amidst concerns about "dis/misinformation" on the platform. YouTube's Community Guidelines, which set out a "COVID-19 misinformation policy" state that the platform "doesn't allow content about COVID-19 that poses a serious risk of egregious harm."[115]

However, in many cases their policy goes well beyond this principle and actively inhibits public discussion about the pandemic or public health responses to the crisis. The policy states:

> *"YouTube doesn't allow content that spreads medical misinformation that contradicts local health authorities' or the World Health Organization's (WHO) medical information*

115 YouTube, COVID-19 Medical Misinformation Policy, https://support.google.com/youtube/answer/9891785?hl=en&hl=en&ref_topic=9282436

about COVID-19. This is limited to content that contradicts WHO or local health authorities' guidance on:

- *Treatment*
- *Prevention*
- *Diagnostic*
- *Transmission"*[116]

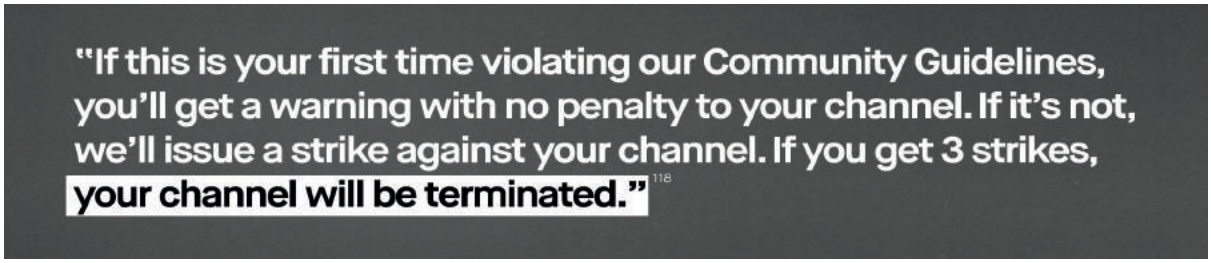
These rules, which govern the world's largest video-sharing platforms, effectively prohibit the ability of users to challenge authorities' guidance on public health measures in response to the virus.

Alongside the policy are examples of types of content which are specifically prohibited on the platform. One such example listed is:

- Content that claims that any group or individual has immunity to the virus or cannot transmit the virus[117]

This could inadvertently mean that content uploaded to the site which suggests that an individual who has been in receipt of a COVID-19 vaccination has immunity to COVID-19 could violate the platform's community guidelines.

The policy states:



YouTube has applied these rules to ordinary users, scientists, and regulated broadcasters

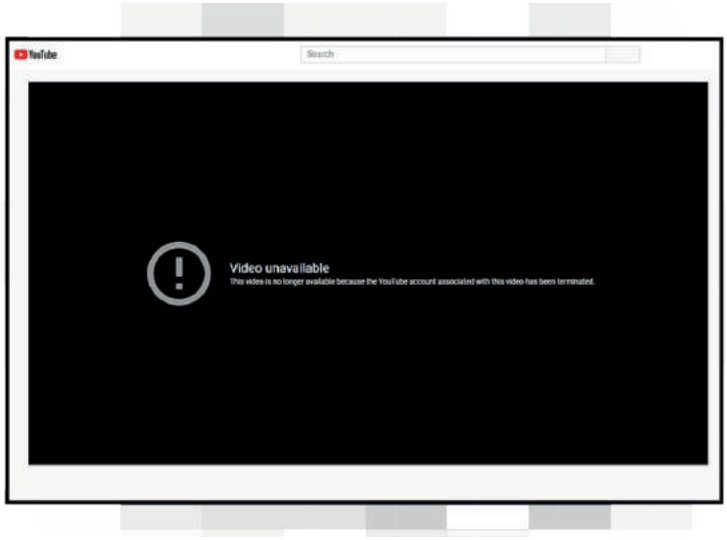
116 116 YouTube, COVID-19 Medical Misinformation Policy, https://support.google.com/youtube/answer/9891785?hl=en&hl=en&ref_topic=9282436

117 Ibid

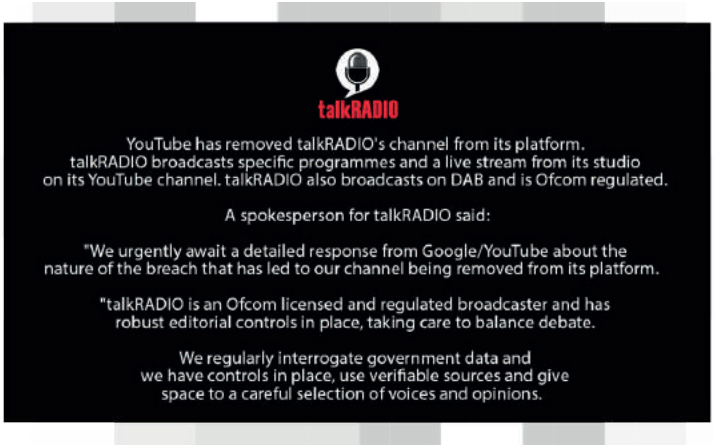
118 Ibid.

alike.

On 5th January 2021, YouTube removed national radio broadcaster talkRADIO from its platform. Despite talkRADIO being an Ofcom-regulated broadcaster, YouTube removed the radio station from its site stating that it had been "terminated for violating YouTube's community guidelines". In a further statement, the platform said "We quickly remove flagged content that violate our community guidelines, including Covid-19 content that explicitly contradict expert consensus from local health authorities or the World Health Organization."[119] talkRADIO, which successfully challenged the decision, had not been investigated by Ofcom for any content related to COVID-19 broadcast on their station.[120]



National radio broadcaster talkRADIO was completely removed from YouTube in early January 2021.



talkRADIO's statement about their banning on YouTube

119 Tobbit,C. Google reinstates Talkradio's Youtube channel after being accused of 'censorship', Press Gazette, Jan 2021, https://pressgazette.co.uk/google-deletes-talkradio-youtube-channel-for-unspecified-violation-of-community-guidelines/

120 Ibid.

Credible reports claimed that talkRADIO's channel was only reinstated after interventions from Government Ministers.[121] From a human rights perspective, this creates a slippery slope. If lawful speech online is only protected at the discretion of politicians, a scenario could develop whereby a government may only choose to sanction the speech of actors they prefer or opinions that they agree with.

The strictness of YouTube's rules leaves little room for free discussion, even around the public policy response to the pandemic. Article 10 of the ECHR states that individuals should have "freedom to hold opinions and to receive and impart information".[122] We find it highly unlikely that qualified limitations on this right could justify the suppression of discussions about public policy in a free and fair democracy.

This example demonstrates the way in which platforms' overzealous content moderation policies are seriously stifling freedom of expression online, with censorship even extending to regulated journalistic content.

YouTube's Community guidelines on content which contradicts "WHO or local health authorities' guidance"[123] has also directly inhibited discussion around the development of potential new treatments for the virus. One such treatment which has been the subject of much discussion by academics, commentators and policymakers has been the anti-parasitic drug Ivermectin. Despite the drug being subject to a number of clinical trials, including most recently at Oxford University, the platform has been quick to censor content related to discussions about Ivermectin.

YouTube's community guidelines specifically prohibits content which not only recommends Ivermectin as a treatment to COVID-19 but also content which includes "Categorical claims that Ivermectin is an effective treatment for COVID-19".[124] This means

121 Tapsfield, J. and Wright J. Google restores TalkRadio's YouTube channel in dramatic U-turn when UK ministers intervened - less than 24 hours after it sparked freedom of speech row by axing station's account 'for airing anti-lockdown views', Daily Mail, 2021, https://www.dailymail.co.uk/news/article-9118263/Covid-UK-Google-restored-TalkRadios-YouTube-channel-U-turn-UK-ministers-intervened.html

122 Equality and Human Rights Commission, Article 10: Freedom of Expression, The Human Rights Act (1998), https://www.equalityhumanrights.com/en/human-rights-act/article-10-freedom-expression

123 COVID-19 medical misinformation policy, YouTube Community Guidelines, https://support.google.com/youtube/answer/9891785?hl=en&ref_topic=9282436

124 COVID-19 medical misinformation policy, YouTube Community Guidelines, https://support.

that various trials and meta-analyses that show an association between Ivermectin and improved recovery are almost impossible to discuss on the platform. As such, YouTube has removed or threatened the removal of large swathes of content related to the drug.

One such instance was the censorship of evolutionary biologist and podcast host Bret Weinstein's YouTube videos on the topic. The platform has taken down several of his videos evaluating evidence with other clinicians and scientists, and issued a "strike" against his channel following discussions about the use of Ivermectin in his Dark Horse podcast.

YouTube's actions contravene a commonly held understanding of the epistemological process which has always been at the heart of free discussion and inquiry in liberal democracies.

google.com/youtube/answer/9891785?hl=en&ref_topic=9282436

"It is a core principle of [redacted] post-Enlightenment democracies that an open [redacted] forum leads to the ongoing discovery of truths. [redacted] [redacted] The Government has abandoned this [redacted] foundational liberal principle. The open forum has been [redacted] recast as a danger to democracy, where the blunt [redacted] tool of suppression is preferred to the forces of [redacted] reason and rationality."

BIG BROTHER WATCH

~~The Role of the State~~

CHAPTER 2: THE ROLE OF THE STATE

Against a backdrop of increasingly censorious tech platforms, the Government has introduced the Online Safety Bill - new regulations for social media companies that will force them to monitor and censor social networks more than ever. The Bill, which is currently undergoing pre-legislative scrutiny, will give Government endorsement to the platforms' restrictive policies, whilst shirking state responsibility for illegal content online.

The Government have designed the plans to explicitly target lawful speech, creating a dichotomy between what can be said online and offline. This sets a seriously dangerous precedent.

Online Safety Bill

The Online Safety Bill follows the 'Online Harms White Paper', which was developed by the Department for Digital, Media, Culture and Sport (DCMS) and the Home Office and published in April 2019. The paper set out proposals to regulate user content across the entire internet. The proposed model centred on imposing a "duty of care" on all companies that enable people to interact with others online, to protect them from "harm". The Government have stated their intention to export this as an international regulatory model.

We believe the Online Safety Bill in its current form is fundamentally flawed and destined to negatively impact fundamental rights to privacy and freedom of expression in the UK. It is our view that placing "duties of care" on online platforms in the name of user safety will force these companies to act as online police and, under the threat of penalties, will force them to over-remove content. We also believe that the Government's approach to this legislation will effectively mean that the legal standard for permissible speech online will be set by platforms' terms of use rather than being clearly set out in statute. Finally, we believe that the broad definition of harm given in the legislation will result in a censorious online environment.

This policy approach is wrong and the legislation must be changed materially in order to protect freedom of expression and privacy.

Duties of care

The legislation increases liability on social media companies for the content on their sites and places new "duties of care" on all "user to user" and "search" services that have "links to the United Kingdom". Excluded from the scope of the legislation are emails, SMS messages, MMS messages, comments and reviews on provider content, one-to-one live aural communications, paid-for advertisements and news publisher content.[125] However, the Government has also reserved the right to extend the duty of care to comments and reviews on provider content as well as one-to-one live aural communications if it is deemed "appropriate" based on the "risk of harm".[126] This could mean the Government could require Zoom calls to be surveilled for "risks of harm" in the future.

The "duty of care" approach is inappropriate for internet intermediaries and other companies online that enable people to interact with one another. It effectively makes companies responsible for how third parties (members of the public) behave towards one another, and puts them in the business of policing people's conversations. To hold companies responsible for the actions of individuals online is to misdiagnose the problem and misplace the responsibility. To effectively apportion general liability and a psychological duty of care to companies over interactions between members of the public could easily make for a very different society to the one we live in today.

According to the Online Safety Bill, regulated user-to-user services must fulfil illegal content risk assessment duties, operational duties with regard to minimising the amount of illegal content or dealing with "harmful" content on the platform; as well as duties which oblige the platforms to "give regard to" freedom of expression, insert reporting and redress mechanisms and fulfil record keeping and review obligations.[127]

125 Draft Online Safety Bill, 2021, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985033/Draft_Online_Safety_Bill_Bookmarked.pdf
126 Ibid.
127 Ibid.

Potentially illegal content

A key operational user safety duty, which applies to all regulated services in scope and is central to the legislation, is set out in clause 9(3) as follows:

> *9 (3) A duty to operate a service using proportionate systems and processes designed to—*
>
> *(a) minimise the presence of priority illegal content;*
>
> *(b) minimise the length of time for which priority illegal content is present;*
>
> *(c) minimise the dissemination of priority illegal content;*
>
> *(d) where the provider is alerted by a person to the presence of any illegal content, or becomes aware of it in any other way, swiftly take down such content.*[128]

Introducing obligations of this nature marks a clear departure from a traditional regulatory approach towards online platforms, held in both the EU and US, which gives platforms immunity from liability for the content on their sites. This principle has been applied in regulatory frameworks with the specific intention of protecting the free expression and privacy of users online. A standard which directly applies is Article 15 of the EU's E-Commerce Directive (this technically still applies to the UK as "EU retained law") prohibits member states from imposing general monitoring obligations on social media companies operating within their jurisdictions.[129]

The duties set out in clause 9 require social media platforms to make judgements on the legality of content and effectively deputise these companies to act as online police. While Cl. 9(3) refers to priority illegal content (that which is specified by the Secretary of State in subsequent regulations), platforms will also be under an obligation to set out in their terms of service how they will "protect users" from all illegal content and to uphold these terms of service consistently.[130] This is particularly problematic when it comes to

128 Draft Online Safety Bill, 2021, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985033/Draft_Online_Safety_Bill_Bookmarked.pdf

129 Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce') https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32000L0031

130 Draft Online Safety Bill, 2021, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985033/Draft_Online_Safety_Bill_Bookmarked.pdf

judgements on the legality of controversial expression.

To take one example of this, the Communications Act (2003) criminalises communications that are deemed to be "grossly offensive"[131]. This legislation has proved to be deeply controversial since it was commenced and has resulted in the criminalisation of speech that merely causes serious offence. In the case of the well-documented "Twitter joke trial", a man was prosecuted after learning that an airport from which he was due to travel was closed due to snow-fall and joking that he would "blow the airport sky high". In Chambers v Director of Public Prosecutions (2012), the High Court overruled the verdict of a magistrate's court that had found the defendant guilty of sending a "menacing electronic communication" under the Communications Act.[132] This demonstrates the complexity of the law in this area and the care that is required when considering the permissibility of speech.

Under their obligations set out in the Online Safety Bill, social media companies would be obliged to set out in their terms of use how they would tackle illegal content such as "grossly offensive" material made illegal by the CA 2003, and uphold these terms "consistently". Under the threat of penalties for non-compliance, this could result in social media companies being overly censorious in their removal of any content which remotely risks crossing this threshold.

The courts, Crown Prosecution Service and the police are all bound by a duty under the Human Rights Act 1998 to act in accordance with the European Convention on Human Rights, including protecting the right to freedom of expression. In practice this can mean that no action is taken against speech that crosses the threshold of "grossly offensive" if this would violate the right free expression. However, no equivalent duty falls upon the platforms in the Online Safety Bill.

The rule of law must be upheld online, but if speech is alleged to cross the threshold into illegality, it should be a matter for the police to investigate and the courts to adjudicate. A swift social media takedown procedure, obligated under the threat of penalties for non-compliance, could inadvertently make users less safe if they are not made aware of potential threats or if potential evidence is speedily removed.

131 Communications Act, 2003, https://www.legislation.gov.uk/ukpga/2003/21/section/127
132 Robin Hood Airport tweet bomb joke man wins case, BBC News, 2012, https://www.bbc.co.uk/news/uk-england-19009344

"Harmful" content

In addition to their duties relating to potentially illegal content, "Category 1 services" (large social media companies) are obliged to fulfil additional duties "to protect adult online safety" including a duty to tackle "content that is harmful to adults". This is set out in the legislation as follows:

> *11 (2) A duty to specify in the terms of service—*
>
> *(a) how priority content that is harmful to adults is to be dealt with by the service (with each such kind of priority content separately covered), and*
>
> *(b) how other content that is harmful to adults, of a kind that has been identified in the most recent adults' risk assessment (if any kind of such content has been identified), is to be dealt with by the service.*
>
> *(3) A duty to ensure that—*
>
> *(a) the terms of service referred to in subsection (2) are clear and accessible, and*
>
> *(b) those terms of service are applied consistently*[133]

These provisions are deeply problematic and pose a serious threat to free speech. The notion of a state-backed system of effectively forcing the removal of online expression which is legal but may cause some form of subjective harm, contravenes accepted human rights standards when it comes to limiting expression. The state should not force the censorship of expression which is lawful and limitations on free speech should be exercised only where necessary, where they are proportionate and where they are clearly prescribed in law.

Further, it is extremely unusual for a private company to be under a "duty of care" to prevent harm resulting from the conduct of others.[134] A duty of care ordinarily refers to a company's duty to ensure its own risk-creating actions do not cause physical injury to others (for example, a stockroom manager has a duty to ensure employees use safe lifting equipment to avoid physical harm). Clearly, these conditions do not apply to internet intermediaries – to provide platforms for people to interact is not a risk-creating action,

133 Draft Online Safety Bill, 2021, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985033/Draft_Online_Safety_Bill_Bookmarked.pdf.

134 See also, UK Supreme Court, Robinson v Chief Constable of West Yorkshire Police, 2018

and there is no risk of physical harm.

Children

Similarly, services which are "likely to be accessed by children" are bound to fulfil further obligations such as preventing children from accessing content that is deemed to be harmful. Set out explicitly, this includes:

> *(3) A duty to operate a service using proportionate systems and processes designed to—*
>
> *(a) prevent children of any age from encountering, by means of the service, primary priority content that is harmful to children;*[135]

Given the huge popularity of social media and the vast number of users on each of the major platforms, the likelihood that a social media site may be accessed by children is high in any case. This means that unless a platform undertakes invasive age verification checks, content moderation on the site in question must be tailored for children.

This directly threatens both free expression and privacy rights online. The measures will force platforms to comply with higher thresholds for deeming content acceptable on their sites unless they verify users' age using ID. This means mandating age verification and would be hugely damaging to privacy rights online. Online anonymity is crucially important to journalists, human rights activists and whistleblowers all over the world. Even tacit attempts to undermine online anonymity here in the UK would set a terrible precedent for authoritarian regimes to follow and would be damaging to human rights globally.

Such a measure would also mean that internet users would have to volunteer even more personal information to the platforms themselves, which would likely be stored in large centralised databases. Further, many people across the UK do not own a form of ID and would directly suffer from digital exclusion.

135 Draft Online Safety Bill, 2021, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985033/Draft_Online_Safety_Bill_Bookmarked.pdf.

The regulator should not seek to enforce companies' terms and conditions

Building on the duties of care, the Online Safety Bill instructs the newly appointed regulator, Ofcom, to draft codes of practice setting out how social media companies can fulfil their obligations when it comes to tackling content which is deemed to be illegal or "legal but harmful". Compliance with the relevant duties is met if a platform takes the steps set out in the codes of practice, which they will have to integrate into their company's "systems and processes".[136]

The effect of this step is to fortify social media companies' terms of use, ensuring that they are upheld, and to clearly identify companies that fail to comply, who risk sanction. Whilst companies consistently upholding their terms and conditions can be seen as a good in and of itself, it is widely recognised that the online data trade means many online companies' terms and conditions are primarily designed for their own economic benefit and legal protection rather than to protect the interests of their users. The terms of service model regulating the relationship between platform and user effectively gives many platforms absolute power and complete discretion as to their application of it.[137] As such, it would seem a controversial position for a Government-appointed regulator to oversee private companies in effectively upholding those terms and conditions – sets of rules that are not neutral, and which have complex implications.

Ensuring companies comply with their terms and conditions raises particularly significant issues where those terms apply to speech issues. Platforms' rules (if not always their enforcement) typically go much further than domestic laws in limiting speech. For example, Facebook's community standards include policies on 'objectionable content' that go far beyond the limits set in domestic law. It would be distinctly wrong for a regulator to oversee the fulfilment of terms and conditions that facilitate the censorship of lawful speech. For the regulator to adhere to and endorse speech standards set in private 'community standards' would show a worrying lack of commitment to the laws and case law on free speech that have evolved in this country over many years. Such proposals would make the Government-appointed regulator complicit in limitations on free speech.

136 Draft Online Safety Bill, 2021, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985033/Draft_Online_Safety_Bill_Bookmarked.pdf.

137 For further analysis, see Digital Constitutionalism: Using the rule of law to evaluate the legitimacy of governance by platforms – N. Suzor, July 2018, in Social Media + Society

Recognising major platforms as the public's modern public squares, Government would do better to play a role in building digital constitutionalism into the tech giants' terms and conditions, that is, embedding human rights and rule of law principles as a basic standard. Government should be using its power and influence to encourage companies to reflect the standards set in UK law and, in particular, human rights law – not to adhere to the standards set in their own self-interested terms and conditions.

The proposals would erode lawful expression online

We agree that the rule of law must be upheld online. However, the Online Safety Bill extends beyond seeking to prevent and prosecute crime and in fact compel companies to prevent "legal harms" online. Following the direction of the Secretary of State, as aforementioned, Ofcom will issue codes of practice specifically about content that is not illegal but that may risk causing "harm" to other users.[138] This, in practice, means that an unelected body will be tasked with delineating what constitutes free speech in the digital environment. This is a dangerous approach to take.

Lawful expression must not be subject to state-sponsored censorship. Any extensions to the limitation on citizens' right to freely express themselves must be decided by Parliament, exercised through statute law, and meet human rights standards. Freedom of expression is protected by multiple human rights frameworks to which the UK is a signatory, notably the European Convention on Human Rights (ECHR). The ECHR is clear that any restrictions on free expression must be "prescribed by law" and necessary in a democratic society[139]. However, the restrictions on free expression that the Government proposes would be prescribed by a regulator, not by law or Parliament. The types of legal "harms" to be regulated will be defined by the Secretary of State via a secondary legislation laid under the negative procedure, taking parliament out of the process. This executive heavy, bureaucratic interference with lawful expression is undemocratic and inherently rights-abusive.

138 Draft Online Safety Bill, 2021, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985033/Draft_Online_Safety_Bill_Bookmarked.pdf.
139 Article 10, Human Rights Act, Equality and Human Rights Commission, https://www.equalityhumanrights.com/en/human-rights-act/article-10-freedom-expression

Loose definitions of harm

The legislation sets out a definition of harmful content (and therefore harm) which category 1 platforms must endeavour to tackle through their terms of use. The definitions set out in clause 46 are:

> *(3) Content is within this subsection if the provider of the service has reasonable grounds to believe that the nature of the content is such that there is a material risk of the content having, or indirectly having, a significant adverse physical or psychological impact on an adult of ordinary sensibilities*
>
> *Or*
>
> *(5) Content is within this subsection if the provider of the service has reasonable grounds to believe that there is a material risk of the fact of the content's dissemination having a significant adverse physical or psychological impact on an adult of ordinary sensibilities, taking into account (in particular)—*
>
> *(a) how many users may be assumed to encounter the content by means of the service, and*
>
> *(b) how easily, quickly and widely content may be disseminated by means of the service.*[140]

In order to protect freedom of expression, restrictions on permissible speech should always be clearly defined in law, to safeguard rights and limit the possibility of overzealous or censorious enforcement. However, a "risk of... having a significant adverse physical or psychological impact" is an overly broad definition and threatens application that would be damaging to free speech. An adverse psychological impact could, for example, refer to a highly offensive joke or footage of an emergency situation. It could even constitute the documentation of social injustice, such as the video of George Floyd's murder which changed debates internationally about race, justice and authority. Such content might cause distress but could be important to see and share for the benefit of society.

140 Draft Online Safety Bill, 2021, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985033/Draft_Online_Safety_Bill_Bookmarked.pdf

The Bill also gives power to the Secretary of State to designate specified categories of "harmful content" that Ofcom must incorporate into its codes of practice.[141] The Government have made it clear that it is not their intention for this list to be static or fixed.

While the Government have set out a definition of "harm" in the Bill, what specific "harms" will be set out in secondary legislation remains opaque. It is also unclear whether these specific "harms" will have to meet the aforementioned standard of posing a risk of having an "adverse physical or psychological impact". Previous statements and documents regarding the legislation have alluded to what these categories could be. The Online Harms White Paper indicated what this could entail by listing a set of initial harms divided into three categories: "Harms with a clear definition", "Harms with a less clear definition", and "Underage exposure to legal content".[142]

The first category consisted of issues covered by the law (e.g. harassment, sale of drugs, modern slavery) and as such represents areas that naturally invite law enforcement intervention and intermediary co-operation, and need not be contextualised in this extra-judicial "harms" model.

The second category set out was deeply problematic and refers to generalised types of content. The items listed include "disinformation", "trolling", "extremist content and activity" and "intimidation".[143]

It is wrong to conflate a policy approach to dealing with unlawful activity online with one dealing with lawful activity simply because it might be deemed 'harmful'. As one online law enforcement expert put it, by the very act of including both lawful and unlawful categories "the authors suggest they are in some way comparable and that they allow for a similar level of debate on their acceptability. They are not and they do not."[144]

Based on previous Government publications and statements in this area, it is possible to consider what categories of "priority content" that is "harmful", might be.

141 Draft Online Safety Bill, 2021, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985033/Draft_Online_Safety_Bill_Bookmarked.pdf
142 Online Harms White Paper – DCMS and The Home Office, April 2019, p.31
143 Online Harms White Paper – DCMS and The Home Office, April 2019, p.31
144 Baines, Dr V. On Online Harms and Folk Devils: Careful Now , Medium, 24th June 2019, https://medium.com/@vicbaines/on-online-harms-and-folk-devils-careful-now-f8b63ee25584

Disinformation

Both disinformation and misinformation were specifically mentioned in the non-exhaustive list of harms in the Online Harms White Paper and with the publication of the Online Safety Bill itself, the Government have indicated that they intend to designate both as categories of "priority content that is harmful to adults". Terms such as "disinformation" can be subjective and easily politicised.

The malleable nature of the concepts of "disinformation" and "misinformation" mean that the threshold for censorship in this area is low. Social media platforms have shown their willingness to make interventions on content that is perceived to be misleading, even by leading academics and elected politicians. The inclusion of "disinformation" as a specified category of content within the Online Safety framework could result in social media companies more frequently arbitrating the speech of academics, pundits and users in general. While disinformation generally constitutes information that is deliberately misleading, it should be reiterated that the legislation is set to cover misinformation too, which is content that is unintentionally misleading or merely inaccurate. This could result in members of the public having any content removed because it is simply considered to be "wrong".

It should generally not be the place of a private company to assess and then instruct their users as to the "reliability" of the information and news sources they access. This is a highly subjective task best fulfilled by internet users themselves, who can optionally conduct wider research or access fact-checking websites online. This is much easier online than it is in a library and offline public spaces. The critical faculties of members of the public are not the responsibility of tech companies. Nor are tech companies best placed to judge the "reliability" of information.

Intimidation

Non-criminal "intimidation" has also been listed as a harm that may be put forward as a priority category of content which is "harmful to adults", but what this would constitute is unclear. Intimidation was referred to in the White Paper in relation to "online abuse

of public figures".[145] However, given the remarkable breadth of the UK's existing laws on speech and communication offences, intervening to repress lawful speech aimed at public figures could threaten political expression and limit democratic participation.[146] It is easy to see how political pressure on the regulator could lead to lawful, albeit unsophisticated or ill-mannered, expression towards politicians and other public figures being suppressed. Such an intervention could unduly limit people's rights and undermine democracy.

Self-harm

Under provisions in the Online Safety Bill, platforms will be under an obligation to minimise content which could risk having an "adverse psychological impact" on users. Therefore, if a post containing a testimonial from someone who has experienced self-harm is deemed to have reached this threshold, it will not only be acceptable for platforms to remove content of this kind but will be expected. As such, the legislation could have the impact of silencing of vulnerable persons' lawful expression. This could have serious impacts on people's rights and mental health.

Whilst the Government have acknowledged that "users should be able to talk online about sensitive topics such as suicide and self-harm,"[147] the legislation will interfere with their right to privately and freely do so in practice. It is common for social media companies to remove the content of those discussing personal experiences of self-harm. As Big Brother Watch has demonstrated, in some circumstances platforms will even remove images uploaded by those who bear self-harm scars, regardless of whether the post in question is related to the topic of self-harm itself. The Bill is likely to deepen this stigmatising activity.

Self-harm: evaluating the case for censorship

145 Online Harms White Paper – DCMS and The Home Office, April 2019, p.24

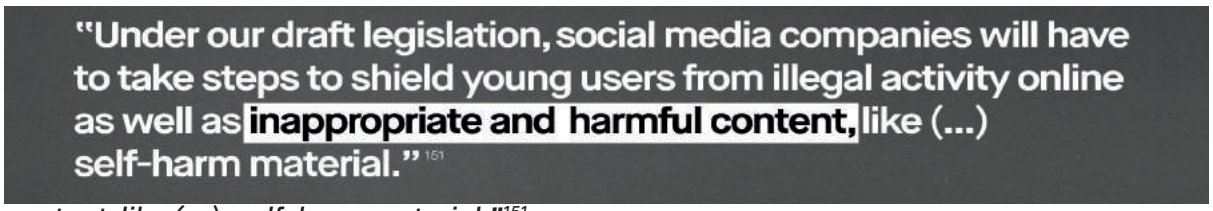
146 See also Written evidence from Index on Censorship (DFF0015) to the Joint Committee on Human Rights' "Democracy, free speech and freedom of association" inquiry, 22 March 2019, http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/human-rights-committee/democracy-free-speech-and-freedom-of-association/written/98529.html

147 Online Harms White Paper – DCMS and The Home Office, April 2019, p.72

Much of the impetus for the Government's "online harms" proposals came from increasing media concern about content relating to self-harm and suicide online. Headlines such as "Instagram 'helped kill my daughter'"(BBC, Jan 2019)[148] have led to an atmosphere of growing fears on this issue.

There is no question that content encouraging or assisting suicide should be removed – encouraging and assisting suicide is an offence,[149] which the CPS is clear applies equally to content online.[150]

However, the Online Safety Bill extends far beyond the law towards an obligation to "shield" internet users from material relating to self-harm. As Secretary of State for



content, like (...) self-harm material."[151]

However, the nature and impact of the "self-harm material" the government refers to is rarely analysed. The Online Harms White Paper, which set out the case for increased online censorship, reported a finding from a study that said 70% of young people who harmed with suicidal intent and 22.5% of young people as a whole "reported self-harm and suicide-related internet use", and reported that this is a "threat".[152] However, this overlooked some key facts from the study. The study found that 3.1% of young people reported exposure to information on how to hurt or kill yourself; the larger statistics include young people who have simply come across news reports online about people who have hurt or killed themselves or who had seen general information about the topic.[153] Therefore, exposure to unlawful and harmful content is much smaller than was otherwise suggested.

148 "Instagram 'helped kill my daughter'" - BBC, 22nd January 2019, https://www.bbc.co.uk/news/av/uk-46966009/instagram-helped-kill-my-daughter

149 Suicide Act 1961, s.2

150 Suicide: Policy for Prosecutions in Respect of Cases of Encouraging or Assisting Suicide, see para.20 p.5, para. 25 p.4, para. 43(11) p.6, – Director of Public Prosecutions, October 2014:

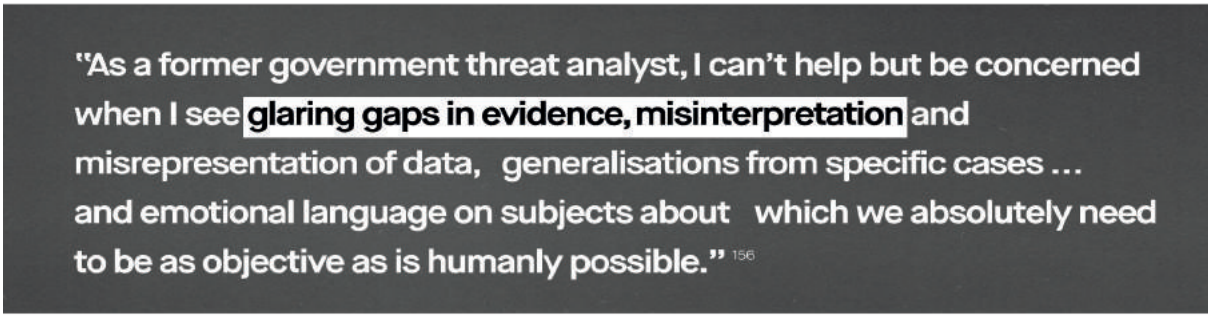
151 Oliver Dowden's Opinion Piece for The Telegraph on the Online Safety Bill – DCMS, 11 May 2021: https://www.gov.uk/government/speeches/oliver-dowdens-opinion-piece-for-the-telegraph-on-the-online-safety-bill

152 Online Harms White Paper – DCMS and The Home Office, April 2019, Box 9, p.19

153 Mars, B et al. (2015). Exposure to, and searching for, information about suicide and self-harm on the internet: Prevalence and predictors in a population based cohort of young adults' Journal of affective disorders, 185, 239-45. Available at: https://doi.org/10.1016/j.jad.2015.06.001

The White Paper also cited the study as having found that 8.2% of young people actively searched for information about self-harm and likewise reported that this is a "threat".[154] However, this overlooked the study's finding that most of this activity was in fact "reassuring": a larger proportion of young people accessed sites offering help, advice and support (8.2%) than sites offering information on how to hurt or kill yourself (3.1%), and most people who had accessed potentially harmful sites had also accessed help sites (81%).[155]

Analysing the White Paper, Dr. Vic Baines, a former law enforcement intelligence analyst, EMEA Trust & Safety Manager at Facebook and now Visiting Associate at the Oxford



Government should not make policy that is built on the misconception that exposure to mental illness is contagious. A shrinking private sphere may deter people from seeking social support and a safe space to freely express themselves. It is important that the internet remains a rich resource for people to openly explore mental health issues, with their rights to privately and freely access information protected.

An excessive increase in executive power

With such a broad definition of harm and an open-ended list of "legal but harmful" areas in scope, we are concerned that the Government can, via ministerial decree, add additional "harms" to this list according to political tides, or content that is merely politically inconvenient to them.

The right to freedom of expression has long been a closely guarded human right, protected in law, with any restrictions subject to full democratic parliamentary process. However, this

154 Online Harms White Paper – DCMS and The Home Office, April 2019, Box 9, p.19

155 Mars, B et al. (2015). Exposure to, and searching for, information about suicide and self-harm on the internet: Prevalence and predictors in a population based cohort of young adults' Journal of affective disorders, 185, 239-45. Available at: https://doi.org/10.1016/j.jad.2015.06.001

156 On Online Harms and Folk Devils: Careful Now by Dr. Vic Baines, Medium, 24th June 2019: https://medium.com/@vicbaines/on-online-harms-and-folk-devils-careful-now-f8b63ee25584

legislation gives a huge amount of power to the executive who will be able to significantly influence the limitations on free speech through Ofcom's codes of practice, and by setting out priority categories of "harmful" content through unamendable secondary legislation, giving direction to the regulator or unilaterally vetoing proposed codes. The codes of practice themselves will also only be subject to secondary legislation.[157]

It is wholly inappropriate for our right to free expression to curtailed by secondary legislation which is unamendable and allows for little Parliamentary oversight. In these circumstances, the power exercised by the online regulator and Secretary of State would bypass the full democratic process, creating a two-tier speech system whereby the increasingly ubiquitous online tier would be, for all intents and purposes, untethered from decades of existing law and highly susceptible to political swings of the day. This situation is precisely what Government should be seeking to prevent – not endorse.

Harsh punishments will encourage companies' zealous censorship

The legislation sets out that a failure on the part of a platform to fulfil their relevant duties of care could result in a fine of up to £18m or 10 per cent of annual global turnover, depending on which is higher. The legislation also lays out a deferred power to impose criminal liability on tech platforms' senior management.

It is unprecedented for the Government to seek to punish technology companies for essentially failing to act as effective law enforcement auxiliaries and even for failing to censor or demote lawful content. If these proposals go ahead, there will be a chilling effect that will motivate companies to monitor, demote and censor expression overzealously. Further, the deferred power to impose criminal liability on companies' senior management sets a terrible example for authoritarian governments around the world to follow, who may justify the imprisonment of social media executives by copying this model.

Such a chilling effect has been seen in Germany, since the Network Enforcement Act 2017 ('NetzDG') was passed. The Government were quick to draw comparisons with the German law and made reference to NetzDG in the White Paper, without acknowledging the serious impact it has had on rights. The Act threatens fines of up to €50 million for

157 Draft Online Safety Bill, 2021, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985033/Draft_Online_Safety_Bill_Bookmarked.pdf

social media companies that fail to remove illegal content within 24 hours. It is extremely heavy-handed and the imposed threat of such a large fine incentivises profit-driven social media companies to err on the side of caution and over-censor content.

Human Rights Watch has called on German lawmakers to "promptly reverse" NetzDG and explained that it is "vague, overbroad, and turns private companies into overzealous censors to avoid steep fines, leaving users with no judicial oversight or right to appeal."[158] Similarly, Article 19 warned that "the Act will severely undermine freedom of expression in Germany, and is already setting a dangerous example to other countries that more vigorously apply criminal provisions to quash dissent and criticism, including against journalists and human rights defenders."[159] The former UN Special Rapporteur on Freedom of Expression, David Kaye, warned that NetzDG "raises serious concerns about freedom of expression and the right to privacy online", and argued that "censorship measures should not be delegated to private entities."[160] The law has also been criticised by the German broadcast media for turning controversial and censored voices into "opinion martyrs".[161] None of these facts have been acknowledged by the Government.

However, in terms of penalties, the legislation also goes much further and gives Ofcom license to seek Service Restriction Orders (e.g. forced removal from the app store) or Access Restriction Orders (ISP blocking), either of which must be approved in court. The proposal for search engine, intermediary and ISP blocking is severe and is a threat to free expression.

These are extremely serious sanctions with wide-ranging effects, including on third parties such as search engines and ISPs and the public more widely. The idea of the British Government appointing a regulator to enforce ISP blocks and search-engine controls over information is extraordinary. Such severe sanctions are chilling and reflect the extreme nature of this proposed legislation which is at odds with fundamental liberal democratic values.

---

158 Germany: Flawed Social Media Law – Human Rights Watch, 14 Feb 2018, https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law

159 Germany: Act to Improve Enforcement of the Law on Social Networks undermines free expression - Article 19, 1 Sept 2017, https://www.article19.org/resources/germany-act-to-improve-enforcement-of-the-law-on-social-networks- undermines-free-expression/

160 Mandate of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, 1 June 2017, https://www.ohchr.org/Documents/Issues/Opinion/Legislation/OL-DEU-1- 2017.pdf

161 Oltermann, P. Tough new German law puts tech firms and free speech in spotlight, The Guardian, 2018, https://www.theguardian.com/world/2018/jan/05/tough-new-german-law-puts-tech-firms-and-free-speech-in-spotlight

This wide range of punishments includes economically and reputationally damaging effects, which would incentivise intermediaries to implement their operational safety duties in an overbearing manner. That is clearly the Government's intention. But this would also mean closer monitoring of users, more intense policing of speech, and a preference for erroneous decisions to demote and censor speech over erroneous decisions to allow speech.

Technological enforcement

A large amount of content moderation is performed exclusively by machines. The Online Safety Bill will support and reinforce this process, owing to sweeping duties to monitor and regulate multiple categories of lawful communications.

AI is a very blunt tool for content moderation, which deals with nuanced areas of speech, law and the adjudication of individuals' rights. Whilst automation can play a role in detecting the most serious illegal material, the Government have previously described building a role for AI in analysing and countering "hate speech" and "to detect and address harmful and undesirable content".[162]

Given greater liability for content and duties of care for users, the Online Safety Bill will force social media companies to ramp up automated content moderation. The effect will be more cumbersome decision-making and a likely increase in unjustified content removals and account suspensions.

The Government have previously used the example of Google's Perspective API – a machine-learning tool that scores the "perceived impact a comment might have on a conversation"[163] to promote this approach to content moderation. Administrators of comment sections (including Disqus, New York Times, El Pais) use Google's API to flag potentially harmful or "toxic" content to moderators. However, this is a crude tool that flags words such as "stupid", "rubbish" and even "racist" as the highest levels of toxicity. The system flags the phrase "I'm angry about women being raped" as the highest level of toxicity, but the phrase "I'm angry at women" is not flagged at all.[164] It is concerning that the Government has lauded this tool without identifying its serious, quite fundamental

162 Online Harms White Paper – DCMS and The Home Office, April 2019, p.80
163 Perspective Website, https://perspectiveapi.com
164 Our own testing on Google's Perspective API: https://perspectiveapi.com

shortcomings. Policymaking in this area should be evidence-led – not led by faith in technology to make complex adjudications about semantics, rights and legal boundaries.

Automation has a limited role – for example, detecting image hashes to identify child sex abuse imagery at scale. However, there is no place for automated mass monitoring or even removal of speech online, nor in complex and nuanced determinations of the legality, much less acceptability, of people's speech online.

Free expression and privacy duties

The legislation sets out a number of additional duties relating to freedom of expression and privacy, protecting "journalistic content" and content which is of "democratic importance". However, far from creating effective protections in these areas, the provisions create points of conflict within the legislation and are, nevertheless, largely outweighed by safety duties that will encourage platforms to over-remove content.

Free expression duties

The duties relating to freedom of expression and privacy, which engage all regulated services, are particularly weak and read as follows:

> *(2) A duty to have regard to the importance of—*
>
> *(a) protecting users' right to freedom of expression within the law, and*
>
> *(b) protecting users from unwarranted infringements of privacy, when deciding on, and implementing, safety policies and procedures..*[165]

Unlike the previously considered operational safety duties, which compel companies to "minimise" illegal or so-called harmful content on their sites, this duty only instructs tech companies to "have regard to the importance" of free expression and privacy.

The duties specifically imposed upon category 1 services are no more conducive to effectively protecting freedom of expression on large social media platforms than the

165 Draft Online Safety Bill, 2021, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985033/Draft_Online_Safety_Bill_Bookmarked.pdf

aforementioned requirements. They compel platforms to undertake impact assessments on the way in which their systems and processes effect freedom of expression and privacy on the platform and set out how they might remedy any threats to these rights on their platform.[166] Once again, this duty is significantly weaker than the operational safety duties and will do little to materially protect free expression online. It also fails to acknowledge that most of the major companies already do limit free expression far beyond that which is prescribed in domestic law, and fails to offer policies to materially remedy this.

The weakness of these provisions was fairly characterised by internet Lawyer Graham Smith when he said:

> *"No obligation to conduct a freedom of expression risk assessment could remove the risk of collateral damage by over-removal. That smacks of faith in the existence of a tech magic wand. Moreover, it does not reflect the uncertainty and subjective judgement inherent in evaluating user content, however great the resources thrown at it."* [167]

The legislation considers the point of conflict between the operational safety duties and duties to protect free expression and privacy in clause 36 (5):

> *36 (5) A provider of a regulated user-to-user service is to be treated as complying with the duty set out in section 12(2) (duty about freedom of expression and privacy) if the provider takes such of the steps described in a code of practice which are recommended for the purposes of compliance with a Chapter 2 safety duty (so far as the steps are relevant to the provider and the service in question) as incorporate safeguards for—*
>
> *(b) the protection of users' right to freedom of expression within the law,*
>
> *or*
>
> *(b) the protection of users from unwarranted infringements of privacy.*[168]

166 Draft Online Safety Bill, 2021, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985033/Draft_Online_Safety_Bill_Bookmarked.pdf

167 Smith, G. Harm Version 3.0: the draft Online Safety Bill, Cyberleagle Blog, May 2021, https://www.cyberleagle.com/2021/05/harm-version-30-draft-online-safety-bill.html

168 Draft Online Safety Bill, 2021, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985033/Draft_Online_Safety_Bill_Bookmarked.pdf

This suggests that so long as platforms follow the code of practice on operational safety duties, this will be sufficient to comply with their free expression and privacy duties. This demonstrates the inherent weakness of the duties to have regard to freedom of expression and privacy, which pay lip service to these fundamental rights in a Bill which otherwise damages them.

Political and journalistic carveouts

In addition to the aforementioned duties, the Online Safety Bill also places an obligation upon category 1 regulated services to protect content of "democratic importance" and "journalistic content". The Government claim that this legislation will not threaten free expression online - however, if this is the case it begs the question of why these carveouts are necessary.

These provisions, clearly borne out of concern that platforms could reprimand politicians in a similar way to former President Trump, include an obligation on platforms to apply the safety duties in a politically neutral manner:

> *13 (3) A duty to ensure that the systems and processes mentioned in subsection (2) apply in the same way to a diversity of political opinion.*[169]

This demonstrates a recognition on the part of the Government that the fortification of and mandated adherence to platforms' terms of use will create a more censorious environment online. However, these provisions effectively exempt politicians themselves from this new system of regulation.

In describing what content of "democratic importance" would constitute, the Bill states:

> *6 (b) the content is or appears to be specifically intended to contribute to democratic political debate in the United Kingdom or a part or area of the United Kingdom.*[170]

The vague nature of this categorisation will only create additional complications for the platforms as they are simultaneously told to remove content which could subjectively be considered "harmful", but not that which is considered a part of "democratic political debate". On the whole, these exemptions present as one rule for politicians, who will

169 Draft Online Safety Bill, 2021, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985033/Draft_Online_Safety_Bill_Bookmarked.pdf.

170 Ibid.

have greater privileges to speak freely online, and one rule for the population at large.

When setting out a duty upon category 1 services to protect "journalistic content", the Bill states that platforms have:

> *A duty ... to make a dedicated and expedited complaints procedure available to a person who considers the content to be journalistic content*[171] *(Where the complainant is the person who shared or created the content in question.)*

Services are also obligated to create such a dedicated and expedited complaints process for all users where action has been taken by the platform on "journalistic content".[172] However, the legislation provides only a loose definition of what "journalistic content" should constitute and states that platforms are to set out a means of identifying journalistic content. The definition given in the Bill is as follows:

> *14 8 (a) the content is—*
>
> *(b) news publisher content in relation to that service, or*
>
> *(ii) regulated content in relation to that service;*
>
> *(b) the content is generated for the purposes of journalism; and*
>
> *(c) the content is UK-linked.*

It is unclear how freelance or citizen journalism would fit within this description. A democratising effect of the internet has been the opening of spaces for marginalised voices, blogs, campaign journalism and more disintermediated news sharing. Citizen journalism online has made a significant contribution to media as a whole, offering new and diverse perspectives, rapid story-telling, inclusive media and audience participation. Citizen journalism has played a major role in 21st Century political events,[173] including the Occupy movement and the Arab Spring, and this has relied on the more equal playing field online for individuals to gain exposure and generate revenue. If carveouts are only afforded to the journalists and media operators that the social media companies choose, this unhealthy monopolisation will only be exacerbated.

171 Draft Online Safety Bill, 2021, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985033/Draft_Online_Safety_Bill_Bookmarked.pdf
172 Ibid.
173 Citizen Journalism, Encyclopaedia Britannica, https://www.britannica.com/topic/citizen-journalism

The Bill would further erode the right to privacy online

The very nature of the legislation, which compels social media companies to take more responsibility and therefore liability for content on their sites means that platforms of this kind will be forced to monitor and surveil users more than ever before. This approach is a serious threat to online privacy and cannot be remedied by asking platforms to simply give "regard" to this fundamental right.

In fact, the Bill will compel social media companies to read the messages of their users to scan for potential "harm". Far from "reigning in" big tech companies, this legislation gives foreign companies license to spy on the communications of British citizens, supporting an exploitative business model that erodes privacy rights.

Mass monitoring and the erosion of privacy on online platforms has been the topic of major controversy in recent years, particularly because companies primarily use it for purposes associated with data profiteering. We should be very cautious of any government proposals that would, in practice, endorse such practices simply because they can serve the secondary purpose of rule enforcement. This is an unacceptable harms trade-off – and companies would be able to justify controversial, profitable data processing practices on the basis that they serve the dual function of protecting users from "harm" online.

In particular the legislation does nothing to protect private spaces online. Many groups rely on the privacy and safety afforded by a private group in order to communicate – particularly those who experience discrimination, are vulnerable or otherwise marginalised. Many people only feel able to express themselves on the basis that their identity, what they are saying and to whom, stays within certain specific circles. This includes marginalised groups, addiction and recovery groups, sexual abuse survivor groups, and community or campaigning groups organising their work. Many of these groups operate on 'public channels' such as Facebook, but the privacy of their groups' visibility, activity and membership can be carefully managed. Imposing monitoring over such groups could have a serious chilling effect.

Private messaging services

The legislation makes clear that private messaging services will be within scope and therefore, platforms will be obliged to uphold duties of care in these channels. This is a dangerous direction and will result in growing surveillance online, even in spaces intended for users to hold a private conversation. The legislation states that:

> *137 (1) "content" means anything communicated by means of an internet service, whether publicly or privately, including written material or messages, oral communications, photographs, videos, visual images, music and data of any description;*[174]

"Safety duties about illegal content" and other obligations to deal with content which is harmful to adults, will therefore extend to private messaging services.

There are important technical issues to consider when imposing the "duty of care" on companies' private messaging channels. Some companies offer structural privacy to their services – for example, the end-to-end encryption offered by instant messaging/VoIP apps WhatsApp and Signal. It is not clear whether the Government's intention is to make privately designed channels of this kind incompatible with platforms' obligations set out in the Bill.

Technology notices

The Bill gives Ofcom the power to mandate the use of technology to identify and remove certain types of illegal content. The legislation states:

> *64 (4) A use of technology notice under this section is a notice relating to a regulated user-to-user service requiring the provider of the service to do either or both of the following—*
>
> *(a) use accredited technology to identify public terrorism content present on the service and to swiftly take down that content (either by means of the technology alone or by means of the technology together with the use of human moderators to review terrorism content identified by the technology);*

174 Draft Online Safety Bill, 2021, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985033/Draft_Online_Safety_Bill_Bookmarked.pdf

(b) use accredited technology to identify CSEA content present on any part of the service (public or private), and to swiftly take down that content (either by means of the technology alone or by means of the technology together with the use of human moderators to review CSEA content identified by the technology).[175]

Given that the Online Safety Bill makes clear that private messaging services are within the scope of the legislation, the provision above implies that certain types of technology could be used to break, erode or undermine the privacy and security provided to messaging services by end-to-end encryption. This could involve the use of a technique known as client-side scanning, which would create vulnerabilities within messaging services for criminals to exploit or could open the door to a greater level of surveillance through use of this technology.[176]

Interference with such companies' technical infrastructure is a matter of great legal and technical debate and would have a profound impact on rights. This does not mean users of such services are beyond the law – law enforcement agencies have a range of powers to seize devices, compel passwords and even covertly hack accounts and devices to circumvent end-to-end encryption.[177] End-to-end encryption does mean that the content of users' communications cannot be subjected to mass monitoring – and given the UK's commitment to upholding human rights and digital security, this should be protected.

The Online Safety Bill should not undermine platform users' ability to have a private conversation online. Private communications are fundamental for our safety and privacy – and are critical for protecting journalists, human rights activists and whistleblowers all around the world. If the Government use this Bill to attack private communications, this will impact upon safety online for all and will set an example for more authoritarian regimes to follow.

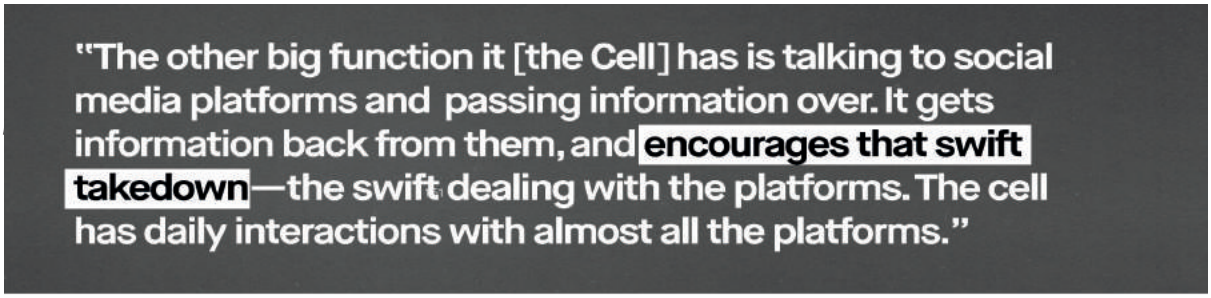
175 Draft Online Safety Bill, 2021, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985033/Draft_Online_Safety_Bill_Bookmarked.pdf
176 Fact Sheet: Client-Side Scanning, The Internet Society, March 2021, https://www.internetsociety.org/resources/doc/2020/fact-sheet-client-side-scanning/
177 See Regulation of Investigatory Powers Act (RIPA) 2000, and Investigatory Powers Act 2016

The Government's "counter disinformation" activity

The Online Safety Bill is being developed during the height of the COVID-19 pandemic where concerns about "disinformation", "fake news" and "antivax" content are high and increased speech suppression is presented as a solution.

On 15th December 2020, during comments in Parliament on plans for the Online Safety Bill, the Secretary of State for Digital Culture, Media and Sport explicitly stated the Government's intention to include so-called "anti-vax" content as a particular "harm" the legislation will seek to stamp out online.[178] How such content could be defined and removed without eroding free expression remains to be seen.

Later that month, the Government appeared to admit recommending the removal of specific pieces of lawful content online. Speaking within a wider discussion about combatting "anti-vax" material online, Sarah Connolly, Director, Security and Online Harms at Department for Digital, Culture, Media and Sport, set out the work of the DCMS Counter Disinformation Cell and the way in which they advise online platforms on content removal. She said:



She was then asked by the Chair of the Committee, Julian Knight:

> *"To be clear, what happens? You see a piece of this, and then send an e-mail and then do an act on it—is that the process?"*

Sarah Connolly replied:

> *"It depends. Each platform is a slightly different set-up. For some of them, we have what is called trusted flagger status. If somebody*

178 Dowden, O. Secretary of State for Digital, Culture, Media and Sport, Statement on Online Harms Consultation, December 2020, https://hansard.parliament.uk/Commons/2020-12-15/debates/1B8FD703-21A5-4E85-B888-FFCC5705D456/OnlineHarmsConsultation#contribution-3303C23B-BC84-43EC-8B98-6408A27A13FF

from the cell says, 'We are worried about this,' that goes immediately to the top of the pile. Whoever it is in whatever company then acts on it."[179]

The Counter Disinformation Cell was initially created amidst concerns about disinformation during the 2019 European Elections,[180] and was later reactivated following the start of the pandemic. The idea that the acceptability of speech online might be policed by a relatively unknown Government unit, without any degree of accountability, is deeply concerning, particularly since the cell in question is requesting the removal of lawful content. The Cell's secretive pressurising on intermediaries to remove unwanted speech, untethered from the rule of law, is at great risk of constituting extrajudicial state censorship. We strongly urge for more transparency over such activities and an immediate stop to Government requests for the removal of lawful communications.

179 Digital, Culture, Media and Sport Committee, Sub-Committee on Online Harms and Disinformation, Oral evidence: Anti-vaccination disinformation, December 2020 https://committees.parliament.uk/oralevidence/1448/html/

180 Ibid.

Recommendations

CHAPTER 3: RECOMMENDATIONS

2021 has already made policymakers sit up and take stock of the overwhelming power that tech platforms have, both over them and over citizens across the world.

Partisanship clouded the debate on Twitter's decision to suspend former President Donald Trump from its platform. However, one question that cut across party lines did prevail through debates around the merits and de-merits of the decision; who controls the censor? If a company or platform has the power to silence the US President, then how that power is held to account is a critical consideration.

Boasting turnover and user bases larger than any single country's GDP or population, the power of these platforms has grown exponentially in recent years. There have been many calls by politicians to "reign them in" without explicitly stating what this could mean. The increasing alignment between state powers and tech giants is not "reigning in" companies, but appropriating their power towards political ends.

We believe the best place to vest power is with people. The role of the state should be limited to upholding the rule of law online, and promoting human rights standards. Big Brother Watch is concerned that the Bill erodes the rule of law in favour of foreign companies' terms and conditions, and undermines human rights standards.

Recommendations for platforms

Platforms must open up their algorithms for public scrutiny

Online platforms are designed with the intention of keeping users on the site for as long as possible. This allows platforms' systems to gather more data on users while keeping users exposed to digital adverts based on information that the site garners about them. This feedback loop itself is protected by amplifying content that catches the user's eye. Algorithmic content ranking promotes posts based on users' preferences which can perpetuate personal biases but also gives more weight to content which is controversial or incendiary.

This is an issue concerning design and not one which can or should be fixed by inhibiting free speech online. Shadowbanning or demoting specific content is de facto censorship

and does little to tackle what is a broader issue. However, a focus on how the design and the processes that platforms use to keep those online hooked is a reasonable consideration.

Algorithmic transparency is crucial, in order to allow policymakers and civic society the opportunity to examine and scrutinise how these vast corporations operate. This kind of transparency should also extend to content moderation and platforms should be willing to publish granular data both on any actions taken on users' posts and on the broader functionality of their algorithms.

In its inception, Twitter presented posts from its users chronologically. Subsequently, the platform has introduced features that promote different types of content based on user preference. However, this system amplifies posts that have generated the most engagement, often prioritising controversy or outrage.

Platforms should consider how content is ranked in a way which does not promote sensationalism. Platforms should also make algorithmic systems open to public scrutiny.

Platforms should reflect human rights principles in their approach to content regulations

When it comes to the permissibility of speech online, major internet intermediaries need digital constitutions that reflect the foundational values of the democracies they serve. This means content policies should reflect human rights principles and avoid limiting expression beyond the limitations of the law.

There is an evolving acknowledgement of the role businesses should play in protecting human rights. In 2008, the UN Human Rights Council in approved the "protect, respect and remedy" framework for business and human rights, resting on three core principles:

1. the state duty to protect against human rights abuses by third parties, including business;

1. the corporate responsibility to respect human rights; and

2. greater access by victims to effective remedy, judicial and non-judicial.[181]

181 UN "Protect, Respect and Remedy" Framework and Guiding Principles, 2008, Business and

This means that whilst States are the primary duty bearers in securing the protection of human rights, corporations have the responsibility to respect human rights - and both entities are joint duty holders in providing effective remedies against rights violations. As a fundamental right this includes a duty to protect freedom of speech.

In 2015, the Internet Governance Forum delivered recommendations on Terms of Service and Human Rights, defining due diligence standards for platforms with regard to three components: privacy, freedom of expression and due process. When considering internet platforms and freedom of expression, those recommendations acknowledged that

> "certain platforms should be seen more as "public spaces" to the extent that occupy an important role in the public sphere"

and that

> "online platforms increasingly play an essential role of speech enablers and pathfinders to information".[182]

Clearly, major online platforms are now among the most widely used public squares. As such, Big Brother Watch believes that major platforms should not censor content beyond the extent to which it would be censored under the law or otherwise prohibited under human rights frameworks. This position is also promoted by the Internet Governance Forum's recommendations of Terms of Service and Human Rights, which say:

> "when platforms offer services which have become essential for the enjoyment of fundamental rights in a given country, they should not restrict content beyond the limits defined by the legitimate law."[183]

As such, platforms should embed human rights principles in their content moderation systems.

Human Rights Resource Centre, https://www.business-humanrights.org/en/un-secretary-generals-special-representative-on-business-human-rights/un-protect-respect-and-remedy-framework-and-guiding-principles

182 RECOMMENDATIONS ON TERMS OF SERVICE & HUMAN RIGHTS, Internet Governance Forum, 2015, https://www.intgovforum.org/cms/documents/igf-meeting/igf-2016/830-dcpr-2015-output-document-1/file

183 Ibid.

Platforms should model enforcement on rule of law principles

Platforms have to yield some power to more democratic forces, because their exercise of power requires limitation if it is to be fair. Currently, the terms of service model effectively gives most platforms absolute power and complete discretion as to their application of it. This needs to change.

We believe that major internet platforms should adopt rule of law principles for enforcement. Government should be promoting rule enforcement that centres transparency of rules, foreseeability of their application, fairness of processes, the right to appeal, and equal and consistent application of the rules.

Government should also work to establish processes for law enforcement to better work with companies online so that policing is not effectively privatised to unaccountable companies in Silicon Valley, but rather is a co-operative process that ultimately protects due process for citizens.

Platforms should ensure that rule of law principles are embedded in their processes and that these are made clear to users.

Expand user controls

Unlike the physical world, users can exercise considerable control of the information and views they are exposed to online by blocking others, muting key words, controlling news feeds, and using age-appropriate controls. User control helps people to mitigate the subjective "harm" they might otherwise be exposed to. We believe companies should work to further expand and simplify user controls over the information they see, the people they are exposed to, and the recommendations they are shown. This approach protects freedom of expression in our online public squares whilst allowing people to create diverse experiences that reflect their own preferences, interests and needs.

Digital literacy, combined with more effective user controls, would allow individuals to take better control of their online experiences.

"To regulate the internet [redacted] is to [redacted] shape the contemporary world and the democratic [redacted] rights we have within it. The Government's proposals for internet regulation will set norms [redacted] for new modes of social [redacted] interaction; inscribe limitations on people's freedom; influence [redacted] power relationships between [redacted] businesses, citizens and the state; and write [redacted] [redacted] enduring rules into a changing world, for millions of people."

BIG BROTHER WATCH

Recommendations for policymakers

Online Safety Bill

The Government's proposals for an Online Safety Bill fundamentally threaten the right to freedom of speech. They must be materially altered or else opposed. We recommend that the Bill is altered in the following ways:

Don't censor lawful speech

Online Safety Bill proposals specifically prohibit certain categories of "legal but harmful" content online. Duties for the restriction of lawful content should be removed from the Bill.

There are already extensive legal restrictions on speech in the UK and the law should be upheld online as it is offline — but it would be dangerous to impose a two-tier system for freedom of expression with extra speech controls for lawful speech just because it is online.

Any further restrictions on our right to free speech must be in line with British law and decided on by a full democratic, parliamentary process — not by a blank cheque handed to authorities and tech companies. The Government's plans to adjudicate which "legal harms" should be within scope by secondary legislation are unsuitable. These legislative mechanisms do not allow for sufficient Parliamentary scrutiny and are not suitable for legislation which has baring on a fundamental human right.

At a minimum, restrictions on lawful speech should be removed entirely from the Bill.

No Silicon Valley Speech Police

The model upon which the Government's Online Safety Bill plans are built would empower and even force big tech companies to conduct mass surveillance of their platforms and act as speech police online. The Bill would also give state-sponsorship to tech companies' own content policies and terms and conditions, which are often far more restrictive on free speech than UK laws. This means we would be even more closely monitored online and controversial or unpopular opinions, marginalised voices and lawful expression would be at greater risk of censorship — with state backing. Far from reining big tech companies in, this Bill would reinforce their power over public discourse.

Introducing huge fines and sanctions on the companies for the hosting of "harmful" content, the Bill would compel content moderators to err on the side of caution and censor more zealously. Further, algorithms are inappropriate for policing speech online and often fail to detect context or nuance resulting in excessive censorship and poor decision-making. Technical enforcement processes are unsuitable mechanisms for presiding over the permissibility of speech. While automation can be useful in detecting serious illegal content such as child sexual abuse material, they should not be used when it comes to general discourse. Deputising Silicon Valley to act as speech police would be a disaster for freedom of expression.

There should be no state-sponsorship of tech companies' censorious Ts & Cs and no encouragement on Silicon Valley to act as speech police.

Protect private conversations online

The Government's proposals also intrude on private communications. Platforms' duty of care will apply to private groups online which are set up to protect those who use them. These spaces are not in the "public domain" and yet will be exposed to greater surveillance from content moderators who will be required to uphold strict rules around the permissibility of speech within them.

The proposals could also mean damaging encryption that protects communications on platforms like WhatsApp and Facebook. These plans could undermine our fundamental right to have a private conversation online.

Private communications are vital for our safety and privacy and are critical to protect journalists, human rights activists and whistleblowers all around the world. The Online Safety Bill's attack on private communications makes people less safe from the serious threats of online crime and surveillance, and sets and a terrible example for more authoritarian regimes to follow.

Private communications should be removed from the scope of the Bill.

Extrajudicial state censorship

The Government's Counter-Disinformation Cell and Rapid Response Unit, based in DCMS and the Cabinet Office respectively, have been tasked with monitoring the timelines of social media users here in the UK, flagging content which is "misleading"[184] with the tech platforms and requesting enforcement action. This is a violation of the right to free expression and due process. Limitations on speech should be set by statute and what speech is permissible should not be at the discretion of secretive Whitehall units. There remains little clarity or transparency around the work of these teams. The British public have a right to know who is responsible for any Government interference with discourse online.

More clarity regarding the structure and operations of these bodies is required.

184 Patel, P. HC, Home Office Questions, Hansard, vol. 689, col. 6, 8 February 2021, https://hansard.parliament.uk/commons/2021-02-08/debates/5F2F0112-3889-4D9A-85E5-019CA14CBD38/Anti-VaccinationExtremism#contribution-ACE7F753-40C9-4995-81AB-946F30F15DFF

THE STATE OF FREE SPEECH ONLINE

BIG BROTHER WATCH