DEMOS

VOTING ONLINE

HOW WOULD THE ONLINE SAFETY BILL AFFECT HARMS ARISING FROM POLITICAL DISCOURSE?

ELLEN JUDSON
VICTORIA BAINES

JULY 2022

This project was supported by Reset.Tech

Reset.

Published by Demos July 2022

15 Whitehall, London, SW1A 2DD
T: 020 3878 3955
hello@demos.co.uk
www.demos.co.uk

CONTENTS

SUMMARY

Self-regulation of social media is coming to an end: governments around the world have decided, for better or for worse, that platforms will need to be held to account for their role contributing to violent attacks, insurrections, and atrocities. The Online Safety Bill has now come to Parliament: it's the UK's attempt to answer the question: how do we reduce risks online, while preserving people's freedom of expression? But in seeking to tackle political harms - election interference, extremism, disinformation, abuse - the question of how to protect political discussion has been one of the more contested elements of the debate around digital regulation.

We examined online conversations around the May 2022 local elections which demonstrates the risks that exist within online political discourse: risks like the spread of harmful narratives through exploitation of platform systems, normalisation of identity-based discrimination, trust in elections being undermined, and the spread of targeted hate attacking people in public life. Under its current form, we do not believe the Online Safety Bill would impact the majority of these risks significantly.

The response to harms exacerbated by online discourse is too often framed as a simple one: find individual pieces of content which themselves cause direct harm, and take them down. Demanding this from platforms will fail to protect freedom of expression: but it also fails those harmed, by seeing harm as something which can only be remedied once it has happened, and not something to be prevented in the first place.

In this report we show that a content-moderation first approach is insufficient, as our findings demonstrate that:

- Online harm exists on a spectrum
- The harm of content cannot and should not be divorced from the context it is in.
- Content moderation is not the sole path to online safety
- Without upstream changes to platform systems, protecting political discourse of this sort can act as a gateway to more serious systemic harms

In this report, we have included examples of content that we analysed. For clarity, we have not taken a view on whether these individual expressions would, for instance, fall foul of a platforms' terms of service (the Bill's metric for platform action against legal but harmful content). Rather, they are intended as illustrations of *systemic harm* - evidence of narratives and tropes which, when encouraged, incentivised, normalised, amplified and scaled through platform systems, cause harm. They have also been bowlderised - edited so that the sense is preserved, but in different exact wording to protect the author's privacy.

INTRODUCTION

Self-regulation of social media is coming to an end: governments around the world have decided, for better or for worse, that platforms will need to be held to account for their role contributing to violent attacks, insurrections, and atrocities. The Online Safety Bill has now come to Parliament: it's the UK's attempt to answer the question: how do we reduce risks online, while preserving people's freedom of expression?

But in seeking to tackle political harms - election interference, extremism, disinformation, abuse - the question of how to protect political discussion has been one of the more contested elements of the debate around digital regulation. For some, the Bill will mandate mass over-moderation of legitimate political content as collateral damage as it seeks to keep people 'safe'.[1]

Others (including Demos) believe that political discourse online could in fact worsen as a result of the Bill - become more prone to abuse, disinformation and hatred, rather than less. In an attempt to protect against the legitimate worries about overmoderation, platforms are being explicitly dissuaded from taking the same action to reduce the risks of harm from the amplification of political and media content. A variety of exemptions, exceptions and exclusions written into the Bill risk meaning that even where platforms are required to have effective and proportionate systems that reduce the risk of harms to users in place, and enforce clear and consistent policies, they will be expected to treat certain kinds of content differently: and the Government is planning to make these exemptions even stronger.

INVESTIGATING CONTENT OF DEMOCRATIC IMPORTANCE

Platforms will have duties to 'protect content of democratic importance'. The 'democratic importance' exception requires that platforms who have duties to specify in their terms of service how they will treat priority forms of legal but harmful content, must also '*operate a service using proportionate systems and processes designed to ensure that the importance of the free expression of content of democratic importance is taken into account*'. Although this remains a vague definition, the framing of 'protecting' this content indicates a likely presumption against removal.

Democratic importance is defined in the Bill as regulated user content or news publisher content which '*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'.*

The Explanatory Notes further set out that *'Examples of such content would be content promoting or opposing government policy and content promoting or opposing a political party.'*

Elections are a cornerstone of a functioning democratic system: but they are under threat. They tend to be periods of heightened party political awareness and discussion. They have historically been fertile grounds for weaponisation and exploitation by bad actors, as they represent unique opportunities to stoke divisions, impact people's behaviour through sharing polarising disinformation, and engage in abuse of political candidates. This behaviour is not only facilitated by platforms, but incentivised and encouraged, and normalised through its amplification - changing the frame of what is deemed 'acceptable political debate' to include identity-based violence, attacks and lies.

As such, in May 2022, we undertook a targeted study of online political discourse around the UK local elections. We identified content which we believe would meet the threshold of 'content of democratic importance', based on its relation to the local elections and discussion of parties and policies. We then analysed the patterns of harm evident in the

1 The Government's response has been that it is not mandating any action to be taken against legal speech online whatsoever, only platforms to enforce their own - privately set - terms of service.

discussion around the elections, and how we believe platforms will be expected to reduce the risk of these harms - or not - once the Online Safety Bill becomes law.

WHAT DID WE DO?

Demos's 2020 *Warring Songs* was a major attempt to move beyond thinking about online harms and disinformation in narrow terms.[2] That year, fake news was the buzzword, but it was clear that information operations were far more diverse in their content than merely false information. Abuse, systems exploitation, information pollution and selective sharing of factual content were all critical tools through which malign actors looked to impact information spaces. The categories developed underpin this paper.

In summary, we sought to identify the following behaviours associated with information operations in the context of the local elections in 2022. Under its current form, we do not believe the Online Safety Bill would impact the majority of these behaviours significantly.

TABLE 1

COMMON HARMS ONLINE IN ELECTORAL POLITICAL DISCOURSE	EXAMPLES OF HARMS ONLINE	IN SCOPE OF THE OSB?
Targeted hate	Doxxing, dogpiling, abuse and disinformation targeting an individual (e.g. a candidate)	**Likely** (if designated priority content) But also likely affected by media exemption, democratic importance and journalistic content exceptions, and exclusion of most paid ads
System Exploitation	Hashtag poisoning, abuse of political advertising, coordinated inauthentic activity	**No**: activity rather than content
Election Interference	Electoral disinformation, political conspiracy theories	**No**: does not cause physical or psychological harm to adults or children
Extremism and discrimination	Identity-based disinformation, violence, abuse and hate	**Likely** (if designated priority content) But also likely will be enabled by media exemption, democratic importance and journalistic content exceptions, and exclusion of most paid ads
Crime	Threats, harassment	**Yes** (Schedule 7)

2 https://demos.co.uk/wp-content/uploads/2019/05/Warring-Songs-final-1.pdf

We examined a snapshot of online political discourse: this does not represent a systematic analysis but demonstrates insights into the nature of political discourse and how certain tropes and harms are reoccurring in the context of an explicitly political debate.

We collected 547,924 tweets via Twitter's API, from 28th April to 9th May 2022, which were returned by the Twitter API for the following phrases or hashtags[3]:

Neutral collection terms (387K Tweets):

- 'Local elections'
- #UKLocalElections
- #LocalElections2022

Political collection terms (115K Tweets):

- #VoteConservative
- #VoteLabour
- #VoteUKIP
- #BackBoris

We also included hashtags known to be associated with political extremism in the UK and from an initial review of Tweets judged likely to contain hateful content, but are not representative of all political affiliations.

#AnneMarieWaters

#ForBritain

#ForBritainMovement

#BritainFirst

#SaveBritain

We then took a random sample of tweets[4] and qualitatively analysed them to establish key themes, trends, and keywords and hashtags relating to the behaviours in our initial matrix, until data saturation was reached. These insights then formed the basis of a deep-dive into other tweets which were using similar keywords and hashtags.

As programmatic access to platforms such as TikTok is limited, we also used a new account to manually search TikTok. Content was searched by keywords such as "local elections" and "#localelections [names of areas in the UK]" as well as for specific figures, groups and topics known to be associated with misinformation likely to be linked to the local elections. While there were examples of extremist content and misinformation, these examples did not relate to the local elections. Content specific to the local elections was aimed at encouraging people to vote or offered legitimate commentary on political events.

We also conducted a manual qualitative review of threads and posts on Mumsnet, identified by searching for 'local elections' and reviewing the first two pages of the first ten threads shared. We were unable to access Nextdoor for research due to location-based account verification requirements.

There were some forms of harm we were interested in analysing but we did not find within in the data we examined; including illegal harms such as criminal threats and harassment, or clear forms of electoral disinformation (such as coordinated activity saying that the election was to be held on a different day).

Readers should be aware that this report contains examples of abusive and offensive language. Content has been bowdlerised to protect privacy - the exact terminology and wording used has been changed to prevent content being discoverable through search, but the sense has been preserved.

3 Note these are case-insensitive, so e.g. 'Local elections' will match 'local elections'

4 Each tweet had a unique ID number - we examined over 700 tweets whose ID number ended in '68' for themes until no new themes were emerging.

RISK ONE
TARGETED HATE

Targeted hate is abuse aimed at specific individuals. These attacks go beyond criticising politicians on the basis of policies or decisions to instead make emotive character attacks which can often reinforce stereotypes when aimed at members of marginalised groups. Sharing private information about an individual (doxxing), many people abusing the same individual at once (dogpiling), individual abusive tweets and personal disinformation are all examples of targeted hate. As well as the possibility of this abuse escalating into physical harm, targeted abuse can have a chilling effect on online participation which is a threat to freedom of speech with victims withdrawing from conversations or witnesses deciding against posting online for fear that something similar could happen to them.

Even within the context of the local elections, this kind of targeted hate in the Twitter dataset was aimed at national figures rather than those standing at a local level. Angela Rayner, the shadow Chancellor, in particular stood out as consistently receiving this kind of targeted hate. Hashtags and abusive nicknames in reference to her were frequently made in the context of discussions about 'Beergate', and therefore had overlap with extremist and conspiratorial harm.

What was notable about the references to Rayner in comparison to other politicians was the use of misogynistic stereotypes of women as sexualised and as liars. The hashtags used built on the contents of a *Mail on Sunday* article that accused Rayner of deliberately attempting to distract Boris Johnson by crossing and uncrossing her legs, and joking about this with other MPs using an offensive slang term. In contrast, nicknames used against Keir Starmer made reference to specific actions, whether it be references to accusations he broke lockdown rules by having a curry in Durham, or perceived political indecisiveness. This is a textbook characteristic of gendered disinformation and abuse campaigns targeting women in public life across the world (as demonstrated in Demos' report *Engendering Hate*); to try to undermine women's credibility, character and question their legitimacy in public life by appealing to sexualised and misogynistic tropes.

"Rayner has been keeping pretty quiet. The Labour gains haven't even prompted a single word…A rabbit caught in head lights feeling guilty #beergate #gingergrowler"

"He supported her when she lied about her hurt feelings about her very own #growlergate tale! Then the few times Boris pushed back, like when he reminded people Starmer had his time as DPP and didn't prosecute Saville, he got attacked by the #ScumMedia and the opposition'

"Has anybody observed that the usual mouthiness of the #gingergrowler has reduced - she's quiet about #BeerStarmer. Maybe she's got another married bloke back to her lair. I bet Keir will regret pushing that Boris should step down"

"@[Journalists] Think you overlooked talking about Flangella's #GrowlerGate and Flip Flop's #durhampartygate. Can you picture them running the country? Exactly why I won't ever vote for Labour!"

"...Pretty sure I asked what it was like being a member of the cult of Flangella and Flip Flop… "

Overall, there were over 400 uses of the hashtag #GrowlerGate, – of #gingergrowler and – mentions of 'flangella'. These accusations, which use slang for a vagina, are designed to be a character attack against Angela Rayner that reinforce harmful misogynistic and sexualised stereotypes of women to undermine her credibility as a politician. The normalisation of these character attacks carry the further risk of discouraging women from choosing to enter politics for fear of facing misogynistic attacks that their male colleagues will not face.

We also found common tropes of gendered disinformation being used in tweets targeting Nadine Dorries, Secretary of State for Digital, Culture, Media and Sport. One tweet made equivalences between Dorries and Carrie Johnson (Boris Johnson's wife): echoing the common misogynistic stereotype of women using sexual relationships to gain political power. Another used aggressive, misogynistic swear words in an abusive tweet directed directly at Dorries, with ageist undertones, along with the hashtag 'ToryScumOut'. Both of these tweets made these comments in relation to the performance of the Conservative party in the local elections.

"Rumours of a reshuffle! Truss will replace Sunak - Nadine Dorries will replace Carrie #LocalElections22"

These discussions clearly highlight the limitations of the Online Safety Bill's approach. The failure of the Bill to mention harms faced by women and girls explicitly has widely been criticised. Even assuming that gendered abuse will be designated a priority harm for which platforms must have clear policies on how they will tackle it, the attacks on Angela Rayner were driven by a story which originated in the media. The media exemption means that platforms are exempt from having to take any action to tackle the risks arising from media content - even when the risk of harm is equivalent to the kinds of user content that would be expected to trigger a platform response. This is not a coherent way to regulate platform systems and processes. How a platform would be expected to mitigate the risks associated with a campaign so clearly originated in and driven by the media has not been clarified.

In our review of Mumsnet discussions around the elections, political discussion covered various topics: there was a lot of discussion of people feeling unable to decide who to vote for, debates about the relationship between supporting a local candidate and supporting a national political party, and whether the local elections was the time to 'send a message'. Amongst this, there was discussion of the actions of politicians and councillors which at times invoked extreme rhetoric against individual politicians or parties:

"Boris Johnson, in my view, through his failure to act on Covid, is the 2nd biggest killer of Brits in Britain since the Nazis"

"If I were a supporter of capital punishment (which I'm not), I would be fine to see the Prime Minister hang".

"The corruption of local councillors is so thoroughly documented - how have there not been arrests? Surprised that they go around without any private security"

"That party are fascist criminals who urgently need to be stopped"

These discussions of politicians are unlikely in themselves constitute incitements to violence. However, were they to be significantly amplified and circulated at scale, there could very well be risks of harm arising from them beyond simply the level of harm or risk posed by the post taken in and of itself.

RISK TWO **SYSTEMS EXPLOITATION**

Systems exploitation refers to bad actors taking advantage of features, or an absence of controls, of online platforms. The ability for anybody to join a conversation, for instance, can be exploited by swamping a conversation with irrelevant information, drowning out authentic discussion. Whether the product of the search engine optimisation (SEO) industry or through shadowy click-farms, content can be shared and amplified inorganically and unauthentically, through deliberate weaponisation or exploitation of known platform systems in order to shape what content and users are promoted, demoted, or moderated. Stopping this means changing the design and functionality of the system, not pursuing types of content: this kind of disinformation is best tackled through increased friction, user controls, bot detection and so on. The OSB does not do enough to demand this.

Tackling this sort of harm is crucial for an effective systems-based approach: which the Online Safety Bill claims to be attempting to do. However, the OSB focuses on the risks of user-generated illegal or harmful content, and fails to foreground how platforms should assess how their systems and processes can be weaponised or exploited so that the harm of a single piece of user generated content is scaled and amplified.

For instance, although platform risk assessments are required to include consideration of systems such as 'functionalities…design and operation of the service', these risk assessments are tied to the risks of specific categories of content which are to be designated in secondary legislation - rather than examining holistically how a platform system may be contributing to overall risks of harm. For content which is harmful to adults (but legal), platforms are required only to state how they plan to treat such content and users who share it in their moderation or curation decisions - rather than changing how their own fundamental systems may be contributing to the harm that users face.

System exploitation is more difficult to identify definitively, as network analysis was out of the scope of this research. However, we did identify instances which appeared indicative of the types of activity and behaviour that seeks to exploit platform systems to promote certain viewpoints or certain kinds of content.

Previous examples have shown the use of uncontroversial hashtags (such as #BackBoris, which has and continues to be widely used in political conversation) to spread extreme ideas, exploiting a common social media feature of allowing users to see or search on popular hashtags.

'17 million people voted for Boris - Labour is over - if you vote for Labour you are voting to support the sexual abuse of children #BackBoris #NeverLabour #Growlergate #RwandaSuccess #DurhamPartyGate #BrexitSuccess #StarmerLies #VoteConservative'

We also saw users posting about the election, drawing attention to their accounts, which hosted far more radical views. In both cases, this shows how the linking capabilities of social media systems can be used to find a new, mainstream audience for extremist views.

For instance, we found discussion of whether the London electoral results were legitimate and whether certain groups should be disenfranchised from voting altogether as they were 'distorting' the results.

'How come we allow nationals of the EU to vote in the local elections, for mayors, assemblies and

councils? Aren't we just adding Remoaners who will distort election results after Brexit?'

Originally tweeted by an account with over 5000 followers, which posted several similar tweets focusing on why EU nationals should be disenfranchised, it was then retweeted by accounts whose other retweets included antivaxx and climate change conspiracy theories.

A similar tweet from the same account claimed that:

'There were meant to be 3 million European Union nationals anticipated to be seeking residency in the UK post-Brexit. After Brexit, they then 'discovered' 6 million. Lots of those people voted in the local elections - they can't in the generals. No surprise that London voted 'locally' how it did then….'

These sentiments were echoed elsewhere: this was retweeted 374 times in our dataset: by the time of our analysis, the account which originally posted it was deleted.

'Obviously Labour will be getting more seats in London. That is due to the fact that there aren't any Londoners living there anymore #LocalElections2022'

Accounts which retweeted one of these tweets include those who in their descriptions include slogans supporting QAnon, anti-vaxx and Covid conspiracy theories, racist conspiracies and various explicitly 'anti-woke' or 'anti-PC' and 'anti-MSM' slogans.

'I love freedom - pro Brexit - I'm awake, not woke - I hate the left - I support Trump - No vaccines, masks or mandates at all.'

This is of particular relevance given the Government's promise that the Online Safety Bill will not lead to 'the last thing we want is for users or journalists to be silenced on the whims of a tech CEO or woke campaigners' - hence the inclusion of elements such as the expedited complaints procedure platforms must give for any action taken against 'journalistic' content. This must be available, as per the Explanatory Notes, 'for users who generate, upload or share what they consider to be journalistic content on the service and creators of journalistic content'. There is a risk that those who consider themselves anti-woke while also promoting Covid disinformation, for instance, will feel they qualify for special exemptions from platform moderation based on the Government's promises.

RISK THREE
ELECTORAL INTERFERENCE

The 'classic' form of electoral interference is where foreign coordinated inauthentic activity creates and amplifies political disinformation online with the intention of interfering with that country's democratic process. However, the spectrum of electoral harms can be much greater: including any disinformation which seeks to sway voting behaviour - whether that is voting for a particular person or party, dissuading or discouraging voting at all, or sharing incorrect information about the voting process that might interfere with people's ability to vote; or misinformation which could have the same results.

In our dataset we did not come across the more 'straightforward' - and more easily fact-checked - forms of electoral interference: such as disinformation claiming the election was on a different day. Instead, we saw a great deal of suspicion and mistrust about the legitimacy of the actions of various institutions and whether the electorate was being manipulated - which lay across a spectrum from understandable concern and legitimate mistrust to more extreme forms of conspiracism.

Conspiracy theories - the notion that powerful institutions and individuals are secretly collaborating to design or engineer specific events at the expense of citizens - undermines trust between citizens, government and other organisations. The most recent high profile example of this is the anti-vaccination movement during the pandemic and the 'stop the steal' movement driven by electoral disinformation in the USA: with drastic and devastating health and political consequences - such as the January 6 insurrection driven by online disinformation and extremism.

Occasional tweets in the dataset called into question the legitimacy of the elections, pointing towards collaboration between the media and political institutions to hide the truth about the election that would render it illegitimate:

"The BBC and the Guardian, and I don't think anyplace else, are still not showing the turnout figures. This is huge - they're concealing the turnout figures, it has to be because they're too low - that lot who got elected have no democratic mandate to govern. If we can get the turnout figures, we can prove the local elections are undemocratic"

However, what was most striking was that the Twitter dataset included significant discussion of two events that were seen as deliberately manipulating the electorate: the publication of the Sue Gray report into reports of breaches of Covid restrictions at Downing Street, and the Durham police's investigation of accusations that Keir Starmer also broke Covid restrictions ('Beergate'). The timings of both of these investigations were repeatedly called into question and framed as being attempts by the civil service or the police to limit the amount of information available to the public and thereby swing the local elections result in favour of the party they were seen as protecting:

"The police in Durham are sitting on hands, until the locals are over - they don't want to hurt the votes for Labour!"

"Who's investigating the police in Durham? Putting off investigating until after local elections - corrupt!"

"Wait a second… All those people who are angry cannot have it both ways - the Party gate findings and £fines were put on a shelf til after elections just for political reasons. And where has that

#SueGreyReport got to?"

General mistrust of the media was also extremely present - including 590 uses of the hashtag #ScumMedia within the dataset.

'Listening to the 3pm news on the BBC R4 - no mention of local elections, of course though, they say Keir Starmer's under police investigation. Pls RT if you agree the BBC are biased!'

'It is shocking they've concluded they will investigate #BeerGate post-elections. Politics, bias, playing games. We know he's broken rules - they know - just obvious. Dreadful from the police in Durham!'

'These right wing Labour spox I'm watching on the news are attempting to spin the bad result in the 2019 elections - which they sabotaged - it's hilarious. No wonder they keep talking about their 'antisemitism' sham. #desperate'

The Online Safety Bill has no way to require platforms to tackle the political harms which can occur at scale from conspiracism - both because the type of harm is out of scope until much farther downstream (when conspiracies become extreme or violent) and because it is overly focused on content moderation - which is not an appropriate response to people expressing political mistrust, in the absence of abuse or dangerous false information. Indeed, by focusing on the need to protect 'democratic' speech, even more than platforms currently do at present, it is likely to tip the balance in favour of allowing conspiracism of this sort to flourish.

Once conspiracies gain traction, it can be very difficult to de-risk them. Conspiracies cannot be effectively fact-checked post by post, at an individual content scale of intervention: fact-checks may even be taken as evidence of the truth of the theory. There are clear conspiratorial narratives that platform systems are vulnerable to amplifying and impacting on people's perception of election results. As a consequence, users exposed to this number of tweets on this subject risk being drawn into further conspiracies and lose trust in a variety of democratic bodies. If information comes to light that proves or disproves these theories, there is no guarantee that it will be as widespread as the original conspiratorial tweets.

Where platform systems interact with this mistrust, to incentivise, encourage, and amplify it, they contribute to creating a community of political discourse which is deeply mired in mistrust - and uses it as an in-group identifier (with hashtags like #ScumMedia being used in user descriptions and not only in posts). In a world where politicians and journalists are not only verbally but physically attacked by those who believe they are bad actors, conspiracism online is not an inconsequential concern.

"We are midway through election results, and though I am sad that we have lost a few of the councils, results aren't as poor as forecasts had suggested. The #ScumMedia failed disastrously to destroy the Tories as they tried to do!"

"It's the #ScumMedia who are the virus".

RISK FOUR
EXTREMISM AND DISCRIMINATION

Identity-based violence, hatred and extremism are known to flourish online. Hateful content and behaviour attacks people on the basis of their gender, race, sexuality, religion, or disability, causing serious direct psychological harm to those targeted by hate - whether targeted as an individual or as a member of a group. Disinformation campaigns weaponise stereotypes and tropes to sow division. The mass shootings in Buffalo, in which a man who became radicalised online by white supremacist websites and message boards killed 10 Black people, is yet another reminder of how extremist attitudes expressed online build narratives that radicalise others with the potential for the most horrific outcomes.[5] After the recent school shooting in Uvalde, false and transphobic claims alleging several different trans women to be the shooter began spreading online, a claim that was repeated by Alex Jones and by a US congressman.[6] And online extremism takes many forms - not all of them straightforward 'illegal content' that can be easily identified and removed.

In our dataset, we identified such identity-based violence and hate occurring within mainstream political discourse. Two specific examples stood out for the way they expressed extremist views in ways that position themselves squarely within the remit of democratic debate. For instance, we found instances of a woman journalist being targeted with antisemitic abuse:

"Why don't you actually do your job: not just be like those other [women] journalists. White Zionist Jews are the only thing Starmer cares about for, and he was bold enough to host the Israel Labour party. Have you asked yet about the #fordereport" [id 86699].

Contained within one tweet is a mixture of antisemitism, misogyny and personal abuse targetting both the individual journalist and a politician. This abusive sentiments are linked with the Forde Inquiry, a delayed report investigating a leaked internal report that discussed the handling of antisemitism complaints within the Labour party.[7] In doing so, this tweet goes beyond purely abusive to instead contribute to a narrative of conspiracy, in which prominent figures in the media are deliberately working to undermine certain politicians and protect others. This narrative is designed to cede mistrust in democracy.

One user tweeted 186 times with claims that supporting the Labour party was equivalent to supporting the sexual abuse of children (the total retweets of tweets in this set was 229, by 32 different users). On its own, while clearly alluding to some extreme view, equating voting behaviour with child abuse was ambiguous. However, further examining the account in question supported this statement with claims echoing common Islamophobic tropes that the Labour party and specific Muslim MPs were involved in covering up a number of high profile child abuse cases such as Rotherham. This kind of content, which clearly relates to opposition of a political party, would be an example of content of 'democratic importance' under the current

5 Sardarizadeh, S. Buffalo shooting: How far-right killers are radicalised online. BBC News. 2022. Available at: https://www.bbc.co.uk/news/blogs-trending-61460468 [Accessed 18/05/2022]

6 Right wing misinformation blames unrelated trans women following Uvalde elementary school massacre, Trans Safety Network. 2022 https://transsafety.network/posts/disinfo-uvalde-shooter/ [Accessed 10/06/2022]

7 The Forde Inquiry. 2022. Available at: https://www.fordeinquiry.org/ [Accessed 18/05/2022]

definitions offered.

This user also used hashtags to accompany their Islamophobic tweet, covering a variety of controversial political topics, including #RwandaSuccess and #DurhamBeerGate, alongside conventional local elections hashtags, such as #VoteConservative. The use of these specific hashtags increases their ability to move into more mainstream and wider election debate: seeking to push voters (perhaps those who are already concerned about topics like 'Beergate') to shape their own voting behaviour based on Islamophobic disinformation.

Much of the discussion on Mumsnet we analysed focused on the issue of trans rights: stories of community members asking their local candidates 'what a woman is', references to the 'Respect My Sex If You Want My X' campaign, and debates about which party to vote for given that with Labour supporting trans rights, many people who had not previously voted Conservative felt 'politically homeless'.

The discussions we analysed employed common narratives that perpetuate and seek to legitimise discrimination against trans people, trans women in particular. These were not restricted to expressions of people's political beliefs on the relationship between sex and gender, but included clear reference to tropes of disinformation which are commonly used against LGBT+ people: such as, attacking the character of individuals who speak out in favour of trans rights, or that women and children are endangered by increased protections for trans people. For instance:

"Look at America - the rights of women are fragile… other political parties appear way too keen to give away our rights and are spineless about protecting children"

"The Tories won't protect sex-based rights: as someone else here said, they've basically made rape legal"

This is the kind of political discourse that the Online Safety Bill a) seeks to tackle and b) in practice is likely to end up protecting. Should disinformation and racist abuse, as is expected, be designated as priority harms, platforms will be required in their terms of service to state how they will treat such content - whether they will take it down once reported, whether they will demote it or demonetise it, what systems they have in place to reduce the risk, for instance, of such a tweet being organically or inauthentically amplified to cause significant harm at scale. However: both of these examples are clearly 'content of democratic importance' as they relate to 'opposing a political party' or policy, and as such platforms might be especially wary of changing any systems that would make harm mitigation more likely - meaning that in practice, far from protecting freedom of expression for everyone, the Bill's current method of protecting 'political' speech is likely to result in reduced online safety for marginalised groups online.

Moreover, transphobic abuse and disinformation is likely *not* to be designated as a priority harm, and transgender rights - such as the inclusion or exclusion of transgender people from the protections from conversion therapy - are certainly within the current definition of content of democratic importance, meaning that transphobic abuse that sought to remove or resist protections for trans people would be expected to be given special regard by platforms when they are making decisions about content moderation or curation.

CONCLUSIONS

Our findings point to three key conclusions:

ONLINE HARM EXISTS ON A SPECTRUM

Trying to draw a fine line between content which definitely falls into specific categories, be that illegal/legal or harmful/not harmful is always going to result in significant error - both by under-moderation, that fails to tackle risks, and over-moderation, that fails to protect freedom of expression. The content that we examined frequently in individual cases, did not cross e.g. an illegality threshold, but demonstrated clear resonances and promotion of narratives which, at scale, are known to cause significant harm.

THE HARM OF CONTENT CANNOT AND SHOULD NOT BE DIVORCED FROM THE CONTEXT IT IS IN.

An individual piece of content, taken alone, may seem extremely harmful - but without attention or amplification, the harm it can do at scale is limited. Likewise, it may seem fairly innocuous - but in using and reinforcing tropes and narratives which are being amplified and replicated across the online information ecosystem, it represents a much greater concern than the same content, physically posted on a flyer in isolation. Locating the harm of content as inherent in the content itself (for which responsibility ultimately rests with the author) is to overlook the ways in which platforms design services driven by commercial, not social or democratic, imperatives, which shape the nature of our political discourse and the risks associated with it.

CONTENT MODERATION IS NOT THE SOLE PATH TO ONLINE SAFETY

The framework set out by the OSB would be most concerned with: a) whether each piece of content we examined was legal, and if so, then b) whether it constituted harm to an adult or child and if so, then c) whether it contravened a platforms' particular terms of service.

We believe that these are the wrong questions to be asking: and in asking them, we are overlooking the trends, the narratives and patterns in political discourse that are systemically reinforced without being collapsible to one individual harmful piece of content: from gendered stereotypes that drive women out of public spaces, to conspiracism that fuels mistrust, to hateful propaganda that stokes division between communities.

And an over-focus on individual content moderation has led to legitimate concerns of overmoderation by platforms, which in turn has led to the democratic importance exceptions to protect political speech.

WITHOUT UPSTREAM CHANGES TO PLATFORM SYSTEMS, PROTECTING POLITICAL DISCOURSE OF THIS SORT CAN ACT AS A GATEWAY TO MORE SERIOUS SYSTEMIC HARMS

However, what is crucial is that all of the examples we include in this report - of disinformation, hate, and abuse - we believe would qualify as 'content of democratic importance', based on their explicit discussion either of political parties, politicians, candidates, elections, or government or party policy. As such, platforms will be under specific duties - not to protect people's free expression *generally* in a particularly robust way - but to ensure they give consideration to how these types of content online will be protected: and implicitly to make different calculations about how their systems should tackle risks associated with this content. That is for the harms which would even be in scope - as discussed, in the OSB as it stands, there is no expectation that platforms take any action to reduce risks associated with media stories prompting gendered disinformation campaigns, or systems exploitation by bad actors.

The Government's response to similar critiques in the past has been that the democratic importance protections are not absolute, and that they mean that platforms must carry out a balancing exercise

which weighs the importance of expressing democratically important content against the risks of harm that could ensue.

'The duty is to ensure that matters concerning the importance of freedom of expression relating to content of democratic importance are taken into account when making decisions. It is not an absolute prohibition on takedown or an absolute protection, but simply something that has to be taken into account.

- Chris Philp, in the Public Bill Committee sessions

Platforms are *already* required to engage in balancing in implementing proportionate and effective systems to tackle risks while respecting rights: so this does not add extra protections by introducing a balancing requirement. This thus indicates that general systemic duties on platforms in the Bill to protect freedom of expression are not truly strong enough to protect freedom of expression. In practice, then, these 'extra protections' will mean platforms are permitted much greater leeway to infringe on users' freedoms to express content that is not tied to discussions of government policy - privileging the power of the government to determine what *qualifies* as political discourse online. Conversely, if the systemic duties are indeed sufficient to protect users, these exemptions serve no extra purpose to protect speech, but could easily incentivise and create space for platforms to be pressured not to act against very real harms where those harms are seen as 'political'.

What this does add, is an expectation of balancing specifically related to the political context of particular instances of content: something which does not make sense within a truly systems-based approach. Experts and platforms themselves have expressed that how this will be expected to be done in practice and at scale is not only a conceptual but a vastly underestimated technical challenge. For instance, one attempt at a 'systems-level' way of approaching this could be for platforms to decide that specific hashtags indicated that the content was related to political debate. That would then give, however, an even clearer pathway through which the system could be weaponised and abused - and the use of hashtag poisoning become a much more successful way to interfere with political discourse.

RECOMMENDATIONS

The Bill must be strengthened in order to adequately tackle these risks. There are two ways this could best be achieved:

IMPROVE HOW THE BILL RESPONDS TO THREATS POSED BY SYSTEMS RATHER THAN CONTENT

- Systems and risk assessments
    - Decouple platforms' risk assessments from being tied to categories of content and require them to assess against kinds of systemic harm
- Transparency
    - Require platforms to publish their risk assessments
    - Require platforms to provide much more comprehensive information through their transparency reports, not limited to content measures but including information on how platform services, systems and processes are designed, tested, modified and deployed
- Include a mechanism through which independent researchers and civil society can have open, consistent and privacy-protecting access to platform data
    - For instance, OFCOM's reporting on data access could lead to a mandatory code of practice for platforms to provide access to data
- Require platforms to conduct and publish electoral risk assessments
- Include provisions to require OFCOM to respond to information incidents and emergencies, and enable independent third parties to raise alerts for emerging incidents (as recommended by Full Fact)

TAKE OUT THE EXEMPTIONS, EXCLUSIONS AND EXCEPTIONS CURRENTLY IN THE BILL - AND STRENGTHEN THE FREE SPEECH PROTECTIONS FOR ALL

As in our joint civil society briefing, we recommend that:

- Schedule 4 should be amended so that the online safety objectives for regulated user-to-user services and regulated search services both include rights protections, such as including that:
    a. A service should be designed and operated in such a way that the human rights, as defined in the Human Rights Act, European Convention on Human Rights and UN Convention on the Rights of the Child, of users and affected persons are protected, including that journalism holds a unique and central role in democratic society, and is recognised in alignment with Article 10 of the ECHR
    b. The amendment should be that in the course of its duties, in carrying out risk assessments, serving information or enforcement notices, and developing Codes of Practice, OFCOM should be required to carry out a rights impact assessment on the systems and risks that they are assessing and the systems or technologies they are recommending (Part 7, Chapter 3).

Strengthening the free speech protections in the bill would allow for it to be amended, removing the media exemption and the exception for content of "democratic importance" and bringing paid ads into scope.

Failing their removal, the following changes would help mitigate the risks of the exemptions, exceptions and exclusions:

- Refine the media exemption to ensure that it does not constitute a 'must-carry', ensuring platforms can take mitigation measures in cases where there is a risk of harm at scale
- Include in any exemptions or special treatment of media or journalists a reference to Article 10 and the necessity of meeting standards of responsible journalism
- Raise the thresholds for who qualifies as a recognised (non-broadcast) news publisher to those which are regulated by an "approved regulator", as defined in the Crime and Courts Act 2013
- Clarify the definitions of political and journalistic speech to be protected in such a way that reduces the likelihood they will privilege the speech of certain users, such as 'content in the public interest' as recommended by the PLS Committee
- Subsume the duty to apply systems equally to a diversity of political opinion within the general duty to have regard to the importance of freedom of expression, including having regard to protecting a diversity of political opinion
- Include an additional duty on platforms to have regard to the importance of preventing any direct or indirect discrimination based on protected characteristics

Licence to publish

Demos – Licence to Publish

The work (as defined below) is provided under the terms of this licence ('licence'). The work is protected by copyright and/or other applicable law. Any use of the work other than as authorized under this licence is prohibited. By exercising any rights to the work provided here, you accept and agree to be bound by the terms of this licence. Demos grants you the rights contained here in consideration of your acceptance of such terms and conditions.

1 Definitions

a 'Collective Work' means a work, such as a periodical issue, anthology or encyclopedia, in which the Work in its entirety in unmodified form, along with a number of other contributions, constituting separate and independent works in themselves, are assembled into a collective whole. A work that constitutes a Collective Work will not be considered a Derivative Work (as defined below) for the purposes of this Licence.

b 'Derivative Work' means a work based upon the Work or upon the Work and other pre-existing works, such as a musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which the Work may be recast, transformed, or adapted, except that a work that constitutes a Collective Work or a translation from English into another language will not be considered a Derivative Work for the purpose of this Licence.

c 'Licensor' means the individual or entity that offers the Work under the terms of this Licence.

d 'Original Author' means the individual or entity who created the Work.

e 'Work' means the copyrightable work of authorship offered under the terms of this Licence.

f 'You' means an individual or entity exercising rights under this Licence who has not previously violated the terms of this Licence with respect to the Work, or who has received express permission from Demos to exercise rights under this Licence despite a previous violation.

2 Fair Use Rights

Nothing in this licence is intended to reduce, limit, or restrict any rights arising from fair use, first sale or other limitations on the exclusive rights of the copyright owner under copyright law or other applicable laws.

3 Licence Grant

Subject to the terms and conditions of this Licence, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) licence to exercise the rights in the Work as stated below:

a to reproduce the Work, to incorporate the Work into one or more Collective Works, and to reproduce the Work as incorporated in the Collective Works;

b to distribute copies or phono-records of, display publicly, perform publicly, and perform publicly by means of a digital audio transmission the Work including as incorporated in Collective Works; The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. All rights not expressly granted by Licensor are hereby reserved.

4 Restrictions

The licence granted in Section 3 above is expressly made subject to and limited by the following restrictions:

a You may distribute, publicly display, publicly perform, or publicly digitally perform the Work only under the terms of this Licence, and You must include a copy of, or the Uniform Resource Identifier for, this Licence with every copy or phono-record of the Work You distribute, publicly display, publicly perform, or publicly digitally perform. You may not offer or impose any terms on the Work that alter or restrict the terms of this Licence or the recipients' exercise of the rights granted hereunder. You may not sublicence the Work. You must keep intact all notices that refer to this Licence and to the disclaimer of warranties. You may not distribute, publicly display, publicly perform, or publicly digitally perform the Work with any technological measures that control access or use of the Work in a manner inconsistent with the terms of this Licence Agreement. The above applies to the Work as incorporated in a Collective Work, but this does not require the Collective Work apart from the Work itself to be made subject to the terms of this Licence. If You create a Collective Work, upon notice from any Licensor You must, to the extent practicable, remove from the Collective Work any reference to such Licensor or the Original Author, as requested.

b You may not exercise any of the rights granted to You in Section 3 above in any manner that is primarily intended for or directed toward commercial advantage or private monetary compensation. The exchange of the Work for other copyrighted works by means of digital file sharing or otherwise shall not be considered to be intended for or directed toward commercial advantage or private monetary compensation, provided there is no payment of any monetary compensation in connection with the exchange of copyrighted works.

c If you distribute, publicly display, publicly perform, or publicly digitally perform the Work or any Collective Works, you must keep intact all copyright notices for the Work and give the Original Author credit reasonable to the medium or means You are utilizing by conveying the name (or pseudonym if applicable) of the Original Author if supplied; the title of the Work if supplied. Such credit may be implemented in any reasonable manner; provided, however, that in the case of a Collective Work, at a minimum such credit will appear where any other comparable authorship credit appears and in a manner at least as prominent as such other comparable authorship credit.

5 Representations, Warranties and Disclaimer

a By offering the Work for public release under this Licence, Licensor represents and warrants that, to the best of Licensor's knowledge after reasonable inquiry:

i Licensor has secured all rights in the Work necessary to grant the licence rights hereunder and to permit the lawful exercise of the rights granted hereunder without You having any obligation to pay any royalties, compulsory licence fees, residuals or any other payments;

ii The Work does not infringe the copyright, trademark, publicity rights, common law rights or any other right of any third party or constitute defamation, invasion of privacy or other tortious injury to any third party.

b Except as expressly stated in this licence or otherwise agreed in writing or required by applicable law, the work is licenced on an 'as is' basis, without warranties of any kind, either express or implied including, without limitation, any warranties regarding the contents or accuracy of the work.

6 Limitation on Liability

Except to the extent required by applicable law, and except for damages arising from liability to a third party resulting from breach of the warranties in section 5, in no event will licensor be liable to you on any legal theory for any special, incidental, consequential, punitive or exemplary damages arising out of this licence or the use of the work, even if licensor has been advised of the possibility of such damages.

7 Termination

a This Licence and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this Licence. Individuals or entities who have received Collective Works from You under this Licence, however, will not have their licences terminated provided such individuals or entities remain in full compliance with those licences. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this Licence.

b Subject to the above terms and conditions, the licence granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different licence terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this Licence (or any other licence that has been, or is required to be, granted under the terms of this Licence), and this Licence will continue in full force and effect unless terminated as stated above.

8 Miscellaneous

a Each time You distribute or publicly digitally perform the Work or a Collective Work, Demos offers to the recipient a licence to the Work on the same terms and conditions as the licence granted to You under this Licence.

b If any provision of this Licence is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this Licence, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.

c No term or provision of this Licence shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.

d This Licence constitutes the entire agreement between the parties with respect to the Work licenced here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This Licence may not be modified without the mutual written agreement of Demos and You.

DEMOS

Demos is a champion of people, ideas and democracy. We bring people together. We bridge divides. We listen and we understand. We are practical about the problems we face, but endlessly optimistic and ambitious about our capacity, together, to overcome them.

At a crossroads in Britain's history, we need ideas for renewal, reconnection and the restoration of hope. Challenges from populism to climate change remain unsolved, and a technological revolution dawns, but the centre of politics has been intellectually paralysed. Demos will change that. We can counter the impossible promises of the political extremes, and challenge despair – by bringing to life an aspirational narrative about the future of Britain that is rooted in the hopes and ambitions of people from across our country.

Demos is an independent, educational charity, registered in England and Wales. (Charity Registration no. 1042046)

Find out more at **www.demos.co.uk**

DEMOS

PUBLISHED BY DEMOS JULY 2022

15 WHITEHALL, LONDON, SW1A 2DD
T: 020 3878 3955
HELLO@DEMOS.CO.UK
WWW.DEMOS.CO.UK