# FRESPECH FORALL

WHY LEGAL BUT HARMFUL CONTENT SHOULD CONTINUE TO BE INCLUDED IN THE ONLINE SAFETY BILL



# FREE SPEECH FOR ALL

Why Legal But Harmful Content Should Continue to Be Included in the Online Safety Bill

A Collaborative Report By:



Reset.













# INTRODUCTION

In the wake of the spike in racist abuse on social media that followed England's loss in the final of the Euro 2020 tournament and the tragic mass shooting in Plymouth there has been a renewed debate about the failings of tech companies and the role of legislation in dealing with the problem of harmful content online. This has come in the midst of the ongoing debates surrounding the forthcoming Online Safety Bill that has been released in draft form and will soon face prelegislative scrutiny.

For years, civil society organisations have done vital work pressuring tech companies to take greater action to reduce harm online. However, it has long been clear that these dangers also require a regulatory solution. The move to introduce internet regulation of this nature is a huge step. As Alex Krasodomski-Jones of Demos has rightly argued, "It is barely an oversimplification to characterise the current debate on internet regulation as a fight over the things people see, and the things they don't." To that end, we encourage debate of the positions outlined in this briefing.

In summary: regulation must demand changes to platform design and processes that disincentivise legal harm. Illegal content should be treated as a policing problem, dealt with forcefully in law and in court and with the use of appropriate detection and enforcement technologies.

Where content violates a platform's terms of service, it must be removed, and the current enforcement gap must be filled. For a platform like Facebook for example, this would include Holocaust denial, vaccine conspiracism and gendered hate.

Legal but harmful content, properly defined and identified by an independent regulator, should be treated as a design problem, and regulation should incentivise platforms to change their systems and processes to reduce it. There is no legal solution, no policing solution, and no content moderation solution, to a problem like disinformation. But there are other solutions, and those solutions should be part of the Bill: changes to platform design and processes.

One of the key areas of debate at present is around the draft legislation's inclusion of socalled 'legal but harmful' content. Clause 46 of the Draft Bill defines "content that is harmful to adults":

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.

Some perceive the inclusion of legal content in this Bill as an untenable threat to freedom of expression. Regulation of our online lives certainly raises legitimate concerns over privacy and freedom of expression. But, as with the offline world, there is a balance of rights to be struck. Government must tackle online harms without infringing on people's freedoms. It must also preserve freedom from abuse and harassment, and protect vulnerable groups who are targets of online hate. We must end up with a form of regulation of platforms that places democratic rights front-and-centre.

Much of the criticism of the inclusion of legal but harmful content within the Bill focuses solely on what harmful speech might be removed by this legislation and ignores the plethora of voices that are already suppressed online due to the often harmful and toxic online environment. If done properly, the inclusion of legal but harmful content within the scope of this legislation could dramatically increase the ability for a wider range of people to exercise their free speech online by increasing the plurality of voices on platforms, especially from minority and persecuted communities. If we are genuinely committed to promoting democratic debate online, preserving the status quo and continuing to exclude these voices is not an option.

This is not just an opportunity to reduce the negative impacts of hostile and prejudiced online behaviour but also a chance to engage in a society-wide discussion about the sort of internet we do want. It is not enough to merely find ways to ban or suppress negative behaviour, we have to find a way to encourage and support positive online cultures. The companies in the scope of the online safety legislation occupy central roles in the public sphere today, providing key forums

through which public debate occurs. It is vital that they ensure that the health of discussions is not undermined by those who spread hate and division.

At present, online speech that causes division and harm is often defended on the basis that to remove it would undermine free speech. In reality, allowing the amplification of such speech only erodes the quality of public debate, and causes harm to the groups such speech targets. This defence, in theory and in practice, minimises free speech overall. This regulation instead should aim to maximise freedom of speech online for more people, especially those currently marginalised and attacked based on their gender, gender identity, race, disability, or sexual orientation, or often, a combination of characteristics.

Those condemning the Bill on free speech grounds underestimate the potential for social inequalities to be reflected in public debate, and disregard the nature and extent of these inequalities in the 'marketplace of ideas'. As such, the position of some 'free speech' advocates can be paradoxical. They claim to be committed to valuing free speech above all else, propagating an unequal debate that further undermines the free speech of those who are already harmed by social inequalities. Some of those currently arguing for the removal of legal but harmful content from this legislation are instead proposing to criminalise speech that is currently legal - a proposal potentially at odds with the aim of preserving free expression.

When content breaks a platform's policies around hate speech it must be removed. However, a regime based on systemic change rather than the takedown of individual pieces of content is more likely to be free speech-preserving. Rather than focusing on whether individuals are increasingly being prosecuted for their speech - in our view a greater threat to freedom of expression than having one's online content removed - the Online Safety Bill focuses on how systems can be adopted that minimise the harm of certain forms of speech. All systems deployed by platforms are active choices: there is no 'neutral' or 'default' position that guarantees complete and open speech for all. The Online Safety Bill offers the potential to shift the scales from amplifying harm to reducing its spread by applying systems and processes to reduce the promotion and targeting of hate and abuse. This will create a more level playing field for all people participating online.

## LEGAL BUT HARMFUL IS NOT NEW

Many of those who have opposed the inclusion of legal but harmful content within the bill have portrayed this as a new and unique 'threat' to freedom of speech. In truth it is neither, and the requirement to mitigate the effect of harmful but legal content is already addressed in existing legislation. Broadcast media, for example, has long had legal obligations in this area.

The 2003 Communications Act placed a duty on Ofcom to set standards for the content of programmes, including "that generally accepted standards are applied to the content of television and radio services so as to provide adequate protection for members of the public from the inclusion in such services of offensive and harmful material". That requirement stemmed from a consensus at the time that broadcasting, by virtue of its universality in virtually every home in the country—and therefore its influence on people's lives—should abide by certain societal standards. The same could be said now about social media, which is even more ubiquitous and, arguably, more influential, especially for young people.

Similarly, in June 2021, the Carnegie Trust published an extensive response to the bill, which highlighted how this principle is also already imposed on some user-to-user platforms via the Communications Act:

Note that the Communications Act already imposes on some user-to-user platforms the obligation to "protecting the general public from videos and audio-visual commercial communications containing relevant harmful material". "Relevant harmful material" includes material containing violence or hatred against a group of persons or a member of a group of persons based on any of the grounds referred to in Article 21 of the Charter of Fundamental Rights of the European Union of 7 December 2000. While some of this material would fall under illegal content, not all the categories are protected by the criminal law. This means that any such types of content would fall to be assessed under 46(3), which might lead to difficulties in the context of lots of low-grade abuse. Arguably, this then constitutes a reduction

in the level of protection. The commitments made in the G7 communique about tackling forms of online gendered abuse will be in part delivered by this clause and to set a strong international lead, the clause needs to be made to work.

Clearly, the inclusion of legal but harmful content in the Online Safety Bill is not a fundamentally new principle and as such does not pose a novel threat to freedom of expression as argued by some opponents.

# NEW POLLING: WHAT DO PEOPLE THINK **ABOUT REGULATION?**

HOPE not hate carried out new polling to explore public attitudes towards harmful content on social media. The findings show clearly that there is widespread concern about extremist and hateful content, and a huge majority believe that such content should not be on social media, even though much of it remains technically legal.

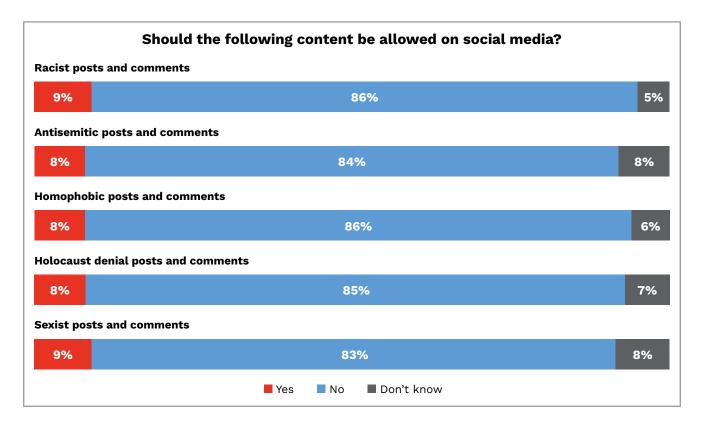
A huge 73% of respondents agreed with the statement: "I am worried by the amount of extremist and hateful content on social media" while 80% agreed with the statement: "I believe in free speech, but there must be limits to stop extremist content and hate speech on social media." Similarly, 73% of respondents agreed with the statement: "Social media companies should be made to remove harmful and hateful content being displayed on their platforms".

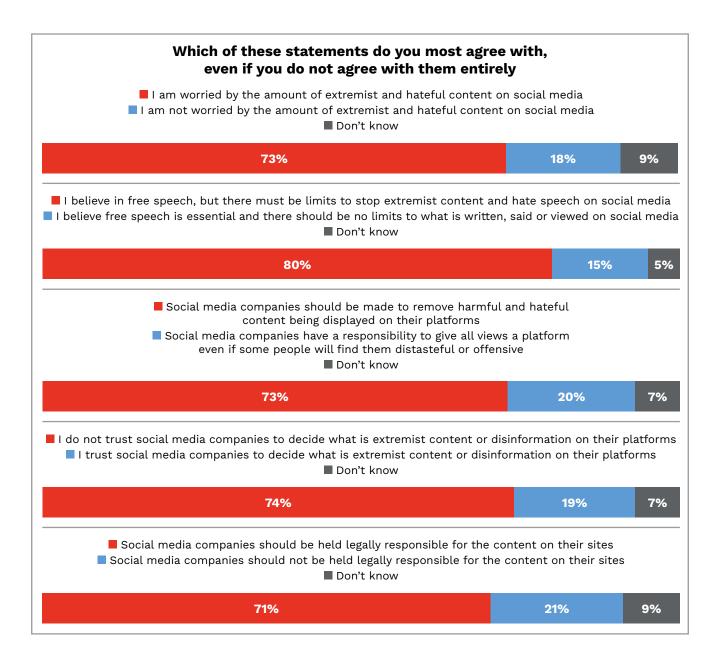
Perhaps unsurprisingly there is a high level of distrust in the ability of social media platforms to deal with this problem themselves, suggesting there is a real public hunger for legislation to tackle the issue. 74% of respondents agreed

with the statement: "I do not trust social media companies to decide what is extremist content or disinformation on their platforms", while 71% agreed with the statement: "Social media companies should be held legally responsible for the content on their sites".

The poll also asked about the types of harmful content online, and whether such content should be permitted, even though much of it is completely legal. In response to the question: "Should the following content be allowed on social media?", 86% answered "no" when it came to racist posts and comments, 84% to antisemitism and 85% to Holocaust denial, 86% to homophobia and 83% to sexist content.

The belief that harmful content should not be allowed on social media is generally accepted across the political spectrum, with almost no difference between 2019 Labour and Conservative voters. There is clearly an overwhelming consensus that hateful content, even when legal, is too present on social media platforms.





#### **METHODOLOGY**

Fieldwork dates: 20-21 July 2021

Sample Size: 1,512

Weighting: UK nationally representative - weighted to age, gender, region, education, and 2019 general

This survey has been conducted using an online interview administered by Focaldata. Focaldata collected data from a representative sample of 1,512 respondents between 20th and 21st July 2021 using its proprietary data collection platform Focaldata Core, which plugs into a global network of panels and uses machine learning to automatically detect and screen out inconsistent and disengaged respondents. Users fill out the surveys in real-time across mobile, desktop, and tablet devices on the Focaldata platform.

## CASE STUDIES

Below are a series of short case studies produced by people concerned about ensuring the inclusion of legal but harmful content within the bill. As noted above, it is important to state that the inclusion of legal but harmful content within the scope of this legislation does in no way make the harmful speech, outlined in the case studies below, illegal. In most instances, it will remain legal to deny the Holocaust, to push Covid-19 conspiracy theories, and to be racist and misogynist. However, by demanding better platforms, with design goals in line with our shared values and processes that empower governments, civil society and a platform's users, this legislation will reduce the spread of, volume of and rewards for harmful behaviour.

#### **COVID-19 CONSPIRACY THEORIES AND ANTISEMITISM**

#### **Danny Stone, Antisemitism Policy Trust**

Anti-Jewish racism is not static. Over millennia, it has evolved and adapted to societal circumstances. Whether it was bubonic plague in the 14th century, tuberculosis in the 19th century or typhus in the 20th century, Jews have been scapegoated by conspiracists and extremists for disease. It was therefore sadly inevitable that Jews would be blamed in some circles for COVID-19 in the 21st century. Antisemitic scapegoating surfaced and spread during the global pandemic, providing ammunition to those seeking someone to blame. Online, memes have been circulating espousing antisemitism, whilst offline, several public figures and others in the public eye have alluded to Jews being the cause of the pandemic. Much of the material, comment and discussion is legal but harmful.

Conspiracy theories that the virus is real and a way for Jews to expand an imaginary web of global influence and profit, or the accusations that the virus is in fact a fake Jewish conspiracy are not necessarily illegal, but they can lead, and have led, to significant isolation of a minority community and real-world harm. Engagement with the aforementioned conspiracy theories easily relates in online spaces to other sinister

narratives, for example that the Jews are the 'real virus' and its primary spreaders, the celebration of Jewish death, or even the so-called 'Holocough', a call to spread the virus to Jews.

Whether it be a caricature of a former French and Jewish health minister pouring poison into a well receiving tens of thousands of social media shares, protesters in Ohio holding up a sign during an April 2020 rally depicting Jews as rats and claiming they are 'the real plague', the refusal of services to visibly identifiable Jewish people for supposed Covid culpability, or the sale of Holocough memorabilia, the consequential harms of antisemitic conspiratorial narratives are manifold. The space for Jews or other minorities targeted by such conspiracy theories is reduced, the tremendous efforts to help people during the pandemic become a greater task for fear of personal safety, the potential threat and requirement to meet it become the burden of Jewish people worldwide, and the online space becomes more toxic and hateful. Legal but harmful material does not exist in a silo, it has consequences, and we have a societal responsibility to protect against this.

#### **FOOTBALL AND RACIST AND** MISOGYNIST ABUSE

#### Hilary Watson, Glitch

The increased prevalence of online abuse in football is receiving unprecedented media coverage, culminating to date in the racist abuse targeting several Black England footballers after the Euro 2020 Final. Women's football is also blighted by online abuse, where Black and minitorised players are racialised, subjected to faith-based and homophobic online abuse, alongside an exorbitant level of sexist abuse that is commonly experienced by women in public life and especially prevalent for sports women. Removing the 'legal but harmful' obligations currently in the draft Online Safety Bill does not move us forward from our experiences now.

The status quo, where online hate is passed off as 'banter' and only a small proportion of incidents of online abuse are followed up by

law enforcement has meant that social media platform self-regulation not only allows legal but harmful content but there is also very little consequence for content that does reach the legal threshold for hate crime and harassment, with so few incidents pursued by the criminal justice system or brought to justice.

The level of online hate experienced by footballers in the UK is not going to go away without increased accountability, transparency and support from social media companies, as well as governments. Several social media boycotts to raise awareness of online abuse, including the four-day social media boycott across professional football in the UK in May 2021, has brought little change.

The current system is not sustainable, where social media companies put the onus on people in the public eye to use features on their platforms to shield themselves from seeing the worst of the abuse directed at them. It should not remain on the platform for every other user to see either.

As we have seen from the fallout of the success of the England football team in the Euro 2020 tournament during a hugely successful campaign where the team is more racially diverse than ever before, social media amplifies hatred but does not provide adequate tools for anti-racism work or allow for meaningful active online bystander intervention - unnuanced automated moderation systems currently in place can mean that talking about abusive behaviour gets would-be activists censored or blocked from the platform.

Offline, Marcus Rashford's wall mural on the streets of Manchester was defaced with racial hatred, yet community response led to a powerful and uniting act of anti-racist action. We must have the opportunities and tools to clean up our online spaces as we do our offline spaces.

Far too many people are being harmed online and platforms are becoming increasingly unsafe. Football boycotts have not brought meaningful, lasting change. The Online Abuse Bill must use a multi-pronged approach to improving the safety of all users through regulations on both illegal content and legal but harmful content. We must ensure social media companies are obliged to make their platforms safe for everyone.

#### **ELECTION DISINFORMATION**

#### **Poppy Wood, Reset**

There are already a number of examples of state backed disinformation campaigns inauthentically amplifying partisan views on Scottish referendum. Twitter and Facebook have identified swathes

of fake accounts linked to the governments of Russia and Iran which amplify messages from bogus pro-independence campaigners. Facebook has also identified, and deleted, a page called Free Scotland 2014 which was traced back to Iran and linked to fake news sites. This is just the tip of the iceberg and it is completely plausible that such efforts to skew electoral processes could result in the ramifications (both offline and online) that we witnessed in the US following the recent Presidential election.

The content shared by these groups and pages was either intentionally misleading (false stories about the Royal family) or intentionally divisive (calling for a change of Conservative party leadership). It is likely that all of the content would have been legal. At present it remains unclear whether election misinformation would be included as a 'harm' in this legislation, but the removal of legal but harmful from this legislation would make it even less likely that it would be dealt with.

#### **ABUSE OF PUBLIC FIGURES**

#### Lexie Kirkconnell-Kawana, IMPRESS

Beyond the scale of abuse experienced by public figures; evidence has been collated on the impact system design has on legal but harmful speech directed at public figures (particularly female politicians) in the Committee for Standards in Public Life 2018 report:

"The first thing we do in the morning is to block and delete online abuse, usually whilst having breakfast. Porridge with one hand, deleting abuse with the other." Office of Rt Hon Diane Abbott MP; Staff of Diane Abbott MP, Individual Oral Evidence, 1 November 2017.

"Social media also bleeds into your 24 hours home life, at night the tweets come in when you're cooking your kids' tea or going to bed. There is little place to hide." Lisa Robillard Webb

That speech may be insulting, sexualised, racialised, imply threat or reveal personal and sensitive information, and while individual instances of this speech may amount to a) malicious communication (criminally liable under MCA 1988) or b) harassment (criminally liable or actionable under the PHA 1997), most of this speech does not reach the threshold for prosecution set out in CPS guidance, and is therefore legal. The purpose of this speech is not to engage in legitimate democratic debate or engage policy decision makers on issues affecting individuals and communities, but to

punish, silence and broadly discourage those likely to be subject to discrimination and abuse from participating in public life. While platforms can engage in moderation and takedown and individual accounts can report instances of this speech, the system design of platforms actively encourages speech patterns which target public figures:

"Extreme positions whether political or moral or abusive, you will get a rise in followers. There is an incentive to go to the extreme." Lionel Barber, Editor of the Financial Times, Individual Oral Evidence, 30 October 2017.

The democratic speech exemption gives the 'bad actor' and the 'poor system design' a legal shield, as all speech - no matter how insulting, offensive, abusive, or objectionable - directed at a public figure could be construed as civic participation, and therefore as democratic content. Should the online safety bill withdraw legal but harmful categories of speech from scope, alongside carving out a protection for democratic content, this could lead to a perverse consequence where platforms are legally disincentivised from moderating speech directed at public figures (particularly political figures), or otherwise designing their systems in such a way that limits or mitigates its impact.

#### ANTI-VAX AND DEEP STATE CONSPIRACY THEORY MISINFORMATION

#### Lexie Kirkconnell-Kawana, IMPRESS

There is no law in the UK that criminalises the publication of inaccurate information (outside of civil liability for reputational inaccuracy, i.e., defamation or malicious falsehood), whether the publication is intentional or accidental. Inaccuracies about innocuous subjects are unlikely to require regulatory intervention, but inaccuracy about matters integral to civic participation and public health and safety have profound effects on social cohesion and daily life. There is plenty of documented evidence of harm that stems from inaccurate information, particularly at scale and enabled by system design. The Institute for Strategic Dialogue in a recent study concerning Germany and Covid-19 information, were able to find 400,000 pieces of false Covid-19 information from more than 1,000 social media users and found that targeted influencing by these actors of the public vaccination debate has been successful in parts of the population.

Similarly, NewsGuard has documented the rise in growth of QAnon groups in the UK, who seek to influence public debate by sowing distrust, targeting politicians and public figures and encouraging the public to withdraw from democratic processes. A series of QAnon Facebook groups appeared in the UK and Ireland starting in April 2020: The Great Awakening - the History of Everything (Cabal. Q. Qanon), a group with 18,200 members was created in April 2020, followed by Q-UK, a group with 2,500 members, created in May 2020. One of the largest UK QAnon Facebook groups (over 9,600), UK Patriot Alliance, was launched in June 2018; its About page stated: "This group is a place to come and learn, wake up the British People and spread the message of #Q." A Twitter account, run by Martin Geddes a London-based technology consultant and photographer who joined Twitter in December 2007, boasted 188,000 followers and since 2018, Geddes has regularly tweeted and retweeted QAnon content.

There is no legal mechanism by which to prevent these actors operating or prevent them from publishing this false information in the UK. Therefore, all information published across these accounts is legal. If the legal but harmful category were to be removed from the Bill, there would be no regulatory incentive for platforms to mitigate the impact of this content as part of their system design.

#### **MEDICAL MISINFORMATION**

#### David Lawrence, HOPE not hate

Few circumstances impress the need for accurate medical information as much as a global pandemic. However, since the onset of COVID-19 there has been an explosion of conspiracy theorydriven medical misinformation online, ranging from claims that 5G technology is causing the health crisis, to the allegation that vaccines are being used to insert microchips into, or to simply kill, recipients. Such scare stories carry potential health implications for both individuals and for society as a whole.

Of course, there is much reasonable discussion to be had about the potential health implications of vaccines, or disputing the efficacy of the current advice from health bodies. There have been many times when a medical product that has been allowed onto the market has subsequently been shown to be harmful, due in part to a lack of transparency.

However, flagrant, conspiracy theory-driven medical misinformation, such as propaganda likening vaccination efforts to Nazi death camps, or alleging that NHS nurses should stand trial for genocide, clearly falls into the category of legal but harmful.

#### **HOLOCAUST DENIAL**

#### Joe Mulhall, HOPE not hate

Since Germany's military defeat in 1945, Holocaust denial has been an attempt by surviving unreconstructed Nazis and their postwar acolytes to whitewash the monstrous crimes of the Third Reich in the hope of rehabilitating the Nazi regime.

Recognising the internet's potential for reaching people at an unprecedented scale, Holocaust deniers were early adopters of online platforms, some as early as the 1980s. Since the 2000s, the spread of social media has had a profound impact, not just on the ability of the denial community to spread their ideas, but more fundamentally on the idea of, and motivation for, Holocaust denial itself.

Holocaust denial is a form of antisemitism and is used to attack and cause harm towards Jews. The idea that Jews have lied about the Holocaust, often, it is argued, to legitimise the State of Israel, has resulted in abuse and attacks against the Jewish community.

While illegal in some parts of Europe, Holocaust denial remains legal in the UK. The inclusion of legal but harmful content within the scope of this legislation would not change that in any way. It would merely ensure that Holocaust denial is not spread on social media platforms where it can cause harm.

#### **GENDERED DISINFORMATION**

#### Ellen Judson, Demos

Gendered disinformation campaigns manipulate information and weaponise gendered stereotypes to discredit people - particularly women in public life - on the basis of their gender and to undermine the fight for gender justice. It is often used as part of a broader political strategy, manipulating existing gendered narratives to silence critics and consolidate power.

Gendered disinformation commonly weaponises stereotypes such as women being devious, stupid, weak, or immoral; or sexualising them to paint them as untrustworthy and unfit to hold positions of power, a public profile or influence within society. Gendered disinformation is a global phenomenon - it is a pattern observed across the world (including in the UK and Europe).1 These campaigns can start online or offline, on social media or in mainstream media, but the possibility for amplification and scale of gendered disinformation campaigns on online platforms makes them a particular cause for concern. Examples of classic gendered disinformation

tactics include faked nude images, videos and sexualised rumours being shared online about women parliamentarians;2 harassment, abuse and threats targeting women journalists<sup>3</sup>; and campaigns playing on tropes of the role of 'mothers' and the 'family' in society to justify curtailing the rights of women and LGBT+ people.4

These campaigns not only take a profound psychological toll on their targets: but threaten their personal safety, with campaigns seeking to justify detention and violence against women in public life and activists for women's and LGBT+

They also pose a wider democratic threat gendered disinformation undermines equal participation in democratic life: seeking to silence existing political critique, trying to prevent opponents being elected, and reducing the space in which it is safe for women and LGBT+ people to be involved in public life.6. This in turn undermines the effectiveness, the equity, and the representativeness of democratic institutions.

There are several factors, however, which mean that gendered disinformation online is likely to fall into the category of 'legal but harmful' rather than illegal speech:

- It is gendered: in the UK, gender is not a category that can define illegal hate speech.
- It operates at scale: individual posts which would not meet a criminal threshold can together form a gendered disinformation campaign. e:
- Though gendered disinformation often weaponises highly abusive language, it is not always so clearly identifiable: it can appear similar to many different forms of legitimate and protected speech:it masquerades as political critique, invokes political beliefs, common views and stereotypes, or can appear similar to counterspeech or satire.7
- Gendered disinformation evolves to take account of the context and environment in which it operates: currently evading algorithmic detection through deliberate misspellings or coded imagery.8 Even if some forms of gendered disinformation became illegal, those who engage in these campaigns would identify where the line was and seek to avoid it, and be able to exploit any ambiguities

As such, relying only on methods to combat individual pieces of illegal content online will be unlikely to greatly reduce the risks associated with gendered disinformation.

#### **ENDNOTES**

- 1 https://eu.boell.org/en/2021/07/09/gendereddisinformation-6-reasons-why-liberal-democracies-needrespond-threat
- https://www.codastory.com/disinformation/howdisinformation-became-a-new-threat-to-women/; https:// www.brookings.edu/techstream/gendered-disinformationis-a-national-security-problem/; https://demos.co.uk/ project/engendering-hate-the-contours-of-state-alignedgendered-disinformation-online/
- 3 https://eu.boell.org/en/2021/07/09/gendereddisinformation-6-reasons-why-liberal-democracies-needrespond-threat
- 4 https://eu.boell.org/en/2021/07/09/gendereddisinformation-6-reasons-why-liberal-democracies-needrespond-threat
- 5 <u>https://eu.boell.org/en/2021/07/09/gendered-</u> disinformation-6-reasons-why-liberal-democraciesneed-respond-threat; https://demos.co.uk/project/ engendering-hate-the-contours-of-state-alignedgendered-disinformation-online/
- https://www.brookings.edu/techstream/gendereddisinformation-is-a-national-security-problem/
- https://demos.co.uk/project/engendering-hate-thecontours-of-state-aligned-gendered-disinformation-online/
- 8 https://www.wilsoncenter.org/publication/maligncreativity-how-gender-sex-and-lies-are-weaponizedagainst-women-online



HOPE not hate Ltd Registered office: Suite 1, 3rd Floor, 11-12 St. James's Square, London SW1Y 4LB, UK