Evidence Submitted to the Equality and Social Justice Committee's Inquiry into Social Cohesion

Professor Matthew Willams

HateLab, Cardiff University
Nisien.ai

Contents

Organisation Description

HateLab is a global hub for data and insight into hate speech and crime. We use data science methods, including ethical forms of AI, to measure and counter the problem of hate both online and offline. The HateLab Dashboard has been developed by academics with policy partners to provide aggregate trends over time and space. Between 2020-2021 the Dashboard was piloted within Welsh Government (Inclusion and Cohesion Communities Division), the National Online Hate Crime Hub (part of the National Police Chiefs' Council), and Galop (a London-based LGBTQ+ charity). HateLab was funded by grants from the Economic and Social Research Council (ESRC), part of UK Research and Innovation (UKRI).

Website: https://hatelab.net/

Nisien.ai is a commercial spin-out of Cardiff University, and owner of all HateLab IP, including the dashboard, which has been rebranded to *Harms Evaluation and Response Observatory (HERO): Detect*. Named after the peacekeeping mythological character in The Mabinogion, Nisien is building a suite of Online Trust & Safety products using advanced AI, to help foster more constructive, factual and healthy online conversations. Nisien has received support from Welsh Government's SMART FIS to build its disruptive *HERO: Resolve* product, that automates the production of de-polarising counter-narratives with Generative AI. It also receives support from the AIRBUS Endeavr Wales Challenge programme to develop its *GenAI Detect* product, which identifies content created and modified by AI, including mis and disinformation. The company completed its seed funding round with the Development Bank of Wales and the British Business Bank (manged by the Foresight Group) in Q4 2024, and currently employs 16 staff based in the Social Science Research Park (sbarc | spark) in Cardiff. To date clients include TikTok, Honda, AIRBUS, Eurovision, EE, British Telecom, Deutsche Telekom, and the Australian Government.

Website: https://nisien.ai

Contact:

Professor Matthew Williams
[redacted]

Context

Responses are provided to the following consultation areas:

- The key issues which impact social cohesion in Wales and consider whether interventions need to target specific groups of people, geographical areas or particular key issues.
- Examples of best practice and other interventions/needed to support social cohesion and overcome tensions. This includes understanding the role the third sector plays in supporting social cohesion and what barriers it faces, including funding.
- What support the Welsh Government provides to community groups and organisations and identify whether there are any opportunities to provide additional support. This includes examining the limitations and barriers which exist given some aspects of support for social cohesion are reserved to the UK Government (for example policing, media and internet regulation), and what action could be taken to overcome these challenges.

Online Hate Speech and divisive mis and disinformation (some of which is created by Generative AI) has received significant government, law enforcement and media attention. The most grievous of these online harms pose risks to the mental and physical health (Williams & Tregidga 2014), increase social division and polarisation (Sunstein 2017) and can lead to extremist activity offline (Williams 2020, Müller & Schwarz 2021).

Defining Online Hate Speech

Cultural and linguistic differences make a general definition of hate speech difficult to formulate. A key problem is the tension between free speech and hate speech, and reaching an agreement on the dividing line across various contexts in order to produce proportionate and effective legislation and operational guidance.

To assist with reaching an agreement, a focus on the impacts of hate speech has been recommended by academics. A set of criteria have been suggested, which if met, should qualify hate speech as criminal (Greenawalt 1989). The speech in question should be criminal if it:

(i) Deeply wounds those targeted
(ii) Causes gross offence to those that hear it, beyond those targeted
(iii) Has a degrading effect on social relationships within any one community
(iv) Provokes a violent response

Others stress the targeted nature of hate speech, where the normatively irrelevant characteristics of individuals or groups single them out. Hate speech then stigmatises victims, who are regarded as ‘legitimate targets’ in the eyes of the perpetrator (Parekh 2012). While these are academic definitions, existing UK legislation and platform policies embody similar criteria.

Legal Definitions of Online Hate

Although many people and organisations use the term 'hate crime', the legal definition in the UK focuses on the word 'hostility', not 'hate'. The police and the Crown Prosecution Service (CPS) have agreed the following definition for identifying and flagging hate crimes:

> *"Any criminal offence which is perceived by the victim or any other person, to be motivated by hostility or prejudice, based on a person's disability or perceived disability; race or perceived race; or religion or perceived religion; or sexual orientation or perceived sexual orientation or transgender identity or perceived transgender identity."*

There is no legal definition of hostility, so the CPS use the everyday understanding of the word which includes: ill-will, spite, contempt, prejudice**,** unfriendliness, antagonism, resentment and dislike.[1]

The CPS does not provide a specific definition of online hate speech, but they do state:

> *"It is more likely that prosecution is required if the offence was motivated by any form of prejudice against the victim's actual or presumed ethnic or national origin, gender, disability, age, religion or belief, sexual orientation or gender identity; or if the suspect targeted or exploited the victim, or demonstrated hostility towards the victim, based on any of those characteristics."* [2]

For a conviction to receive enhanced sentencing in court, there needs to be sufficient evidence to prove the hostility element. In short, the law treats hostility as an aggravating feature when it is:

- linked to a criminal offence
- in some way about one of the protected characteristics

The relevant provisions within the Crime and Disorder Act 1998 and Criminal Justice Act 2003 use the same terminology in setting out aggravation:

- at the time of committing the offence or immediately before or after doing so, the offender *demonstrated* towards the victim hostility based on the victim's membership (or presumed membership) of a specified group(s); or
- the offence was *motivated* (wholly or partly) by hostility towards members of a protected characteristic based on their membership (or presumed membership) of that specified groups(s).

[1] https://www.cps.gov.uk/crime-info/hate-crime
[2] https://www.cps.gov.uk/legal-guidance/social-media-and-other-electronic-communications

Stirring up racial and religious hatred, and hatred based on sexual orientation, are offences under the Public Order Act 1986. Stirring up racial hatred is committed when someone says or does something which is *threatening, abusive or insulting*, and the person either intends to stir up racial hatred or makes it likely that racial hatred will be stirred up. Stirring up religious hatred or hatred on the grounds of sexual orientation is committed if a person uses *threatening words or behaviour* and intends to stir up hatred.[3]

Section 127 of The Communications Act 2003 covers the sending of improper messages via a public information system. Section 127(1)(a) states an offence is made if a message is sent that is grossly offensive, indecent, obscene, menacing or false. The Malicious Communications Act 1988, section 1, deals with the sending to another of any article which is indecent or grossly offensive, or which conveys a threat, or which is false, provided there is an intent to cause distress or anxiety to the recipient.

In practical terms, hateful social media posts (other than those which amount to specific offences in their own right, such as making threats to kill, blackmail, harassment etc.) are considered criminal if:

- Their content is grossly offensive[4]
- Their content is threatening *or* abusive *or* insulting *and* is intended to *or* likely to stir up racial hatred
- Their content is threatening *and* is intended to stir up hatred on the grounds of religion or sexual orientation

It is important to note that when considering cases involving hateful communications, prosecutors operate a high threshold at the evidential stage and consider whether a prosecution is in the public interest based on the nature of the communication and the impact upon the targeted victim. They must also be satisfied that the communication is not protected under the free speech principle (under Article 10) of the European Convention on Human Rights, that provides the freedom to cause offence.[5]

[3] Note that the threshold is higher for the latter set of offences: "threatening words or behaviour" versus "threatening, abusive or insulting". Further, only the latter set of offences contain an express freedom of expression clause to balance the right to free speech with the duty of the state to protect the rights of others and to act proportionately in the interests of public safety to prevent disorder and crime (although Article 10 is relevant to all offences). Stirring up hatred means more than just causing hatred, and is not the same as stirring up tension. It must amount to hatred against a whole group – rather than hostility to just one person – and manifest itself in such a way that public order might be affected.

[4] The test for "grossly offensive" was stated by the House of Lords in DPP v Collins to be whether the message would cause gross offence to those to whom it relates (in that case ethnic minorities), who need not be the recipients. The case also confirms that it is justifiable under ECHR Art 10(2) to prosecute somebody who has used the public telecommunications system to leave racist messages. The European Commission has held that extreme racist speech is outside the protection of Article 10 because of its potential to undermine public order and the rights of the targeted minority: Kuhnen v Germany 56 RR 205. Prosecutors must be satisfied that a prosecution is required in the public interest and that, where Article 10 is engaged, on the facts and merits of the particular case it has convincingly been established that a prosecution is necessary and proportionate. Particular care must be taken where a prosecution is contemplated for the way in which a person has expressed themselves on social media (Williams and Mishcon de Reya 2019).

[5] Speech that crosses the threshold must be more than shocking or distasteful. Incitement offences require the consent of the Attorney General in order to bring charges.

Platform Definitions

Most of the large social media platforms have policies that inform the moderation of content posted by users, including content that targets individuals based on perceived group membership or association.

Generally, platform policies afford protections based on sexual orientation, sex, gender, gender identity, race, religion or belief, ethnicity, national origin (and nationality), disability, disease, caste and age.

Online hate includes but is not limited to:

- Derogation (contempt, stereotyping and negative generalisations, hateful slurs and epithets, and delegitimisation)
- Dehumanisation (comparing individuals and groups to, or labelling them as non-human entities)
- Adversarial elevation (describing groups as threateningly superhuman, emotionless)
- Threatening harm (threatening violent harm, non-violent harm, exclusion)
- Inciting harm (inciting violent harm, non-violent harm, exclusion)
- Glorification of hateful entities and denial of atrocities
- Animosity (undermining groups, accusation of special treatment, mocking and insensitive humour, othering)
- In-group superiority (claiming superiority over an out-group)

At the time of writing, X's (formerly Twitter) Hateful Conduct policy stated:

> *"You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories."* [6]

Examples of what X do not tolerate includes, but is not limited to, behaviour that harasses individuals or groups of people with:

- Violent threats
- Wishes for the physical harm, death, or infection with disease of individuals or groups
- References to mass murder, violent events, or specific means of violence in which/with which such groups have been the primary targets or victims
- Behaviour that incites fear about a protected group
- Repeated and/or non-consensual slurs, epithets, racist and sexist tropes, or other content that dehumanises, degrades or reinforces negative stereotypes

[6] https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

X also forbids the use of hateful imagery, such as symbols associated with hate groups, images depicting others as non-human, and moving images (e.g. gifs, video) containing any of the above.[7]

Prevalence of Online Hate Speech

Research shows the internet has been used to spread socially divisive ideology, including hate speech, since its inception. While research within government departments, NGOs, academic institutions and corporations has attempted to measure and monitor online hate speech, the data-landscape remains patchy due to technical, conceptual and definitional issues that are yet to be fully resolved. Presented below is a summary of the latest estimation efforts across various sectors.

Information from Government Commissioned Surveys

Statistically representative government surveys find similar levels of prevalence:

- Ofcom's Adults' Media Use and Attitudes Survey, conducted in 2018, found **53 per cent** of internet users had viewed hateful content in the last year (Ofcom 2019).[8]
- Their 2020 survey of 12-15 year-olds found that **half had encountered hateful content online**, an increase on the 2016 figure of **34 per cent** (Ofcom 2021).
- The 2022 Ofcom UK Online Experiences Tracker, that asks respondents what they experienced or encountered online in the last 4 weeks, found **11 per cent** of respondents had personally seen or experienced hateful, offensive or discriminatory content that targets a group or person based on specific characteristics like race, religion, disability, sexual orientation or gender identity (Ofcom 2022a).
- Experiences varied significantly by ethnicity of respondent, with **22 per cent** of Mixed Ethnic, **16 per cent** of Asian and **18 per cent** of Black respondents reporting they had seen or experienced these harms, compared to **9 per cent** of whites.
- Hateful behaviours were second most likely to bother or offend users, after animal cruelty.
- Black users were more likely to report being *really* bothered after their experience compared to white users (**38 per cent** compared to **18 per cent**).
- Social media was more likely than average to be the type of technology in use when users encountered hateful, offensive or discriminatory content (**64 per cent**) (Ofcom 2022b).

[7] Recent controversial changes to Platform hate speech policies include the removal of 'misgendering' and 'deadnaming' protections in the case of trans individuals and the removal of protections for lesbian, gay and bisexual individuals.
[8] Questions on encountering hateful contend were not included in the 2020 - 2022 reports.

Information from Police Crime Statistics

Police data on online hate crime has previously been published, but it is flawed due to inconsistent recording across forces.

- Experimental statistics published by the Home Office for 2017/18 showed hate crimes with an online element amounted to **1,605 in the period** (2 per cent of all hate crime recorded).
- Race hate crime accounted for **52 per cent** of all online cases in 17/18, sexual orientation **20 per cent**, disability **13 per cent** and transgender identity **4 per cent**.
- Violence against the person accounted for **80 per cent** of these online hate crimes, while public order offences accounted for **14 per cent**, and other notifiable offences **6 per cent**.
- Summaries have not been published since 2017/18 due to concerns over the accuracy of the data supplied by police forces.
- Crown Prosecution Service statistics for the same period show that there were **435 offences** related to online hate under section 127 of the Communications Act 2003 and section 1 of the Malicious Communications Act 1998, a **13 per cent** increase on the **386** recorded in the previous year.[9]

Information from Platform Detections and Removals

Some platform statistics on the detection and removal of content that violates terms of use are also publicly available, but the percentage of posts that meet the required threshold for moderation set by platforms is very low compared to overall traffic (see section on counter-measures for a discussion of platform transparency reports).

- Vidgen et al. (2019) found that the percentage of posts removed for being hateful across four of the largest platforms was around **0.001** in 2018 (0.001 per cent on Facebook and YouTube, 0.0001 per cent on Reddit, and 0.2–0.3 per cent on Twitter).[10]
- To maximise profit, most platforms will balance moderating some extreme content, thus losing some free-speech absolutist users, while gaining new users because of a 'safer' platform for minorities and the vulnerable. As most large platforms generate significant amounts of profit via advertising, they are more likely to moderate content to attract more users and the widest array of clients.
- However, because of subpar moderation technology, all large platforms routinely mislabel moderate content (known as false positives) and remove non-significant amounts by mistake. Therefore, it is not feasible to judge the overall extremity of content on a platform by examining moderation statistics (Liu et al. 2022).

[9] More recent statistics on online hate crimes and their prosecution have not been provided by the Crown Prosecution Service in its latest hate crime reports.
[10] The platforms do not segment by targeted characteristic or country of origin.

Information from Academic Research

Academic survey studies of prevalence have found high levels of exposure to online hatred.

- A large representative survey of 15-30 year-olds, covering the US, UK, Germany and Finland, found on average **43 per cent** had encountered hate material online. This rose to just over half of those surveyed in the US, while **39 per cent** of UK respondents reported encountering hateful or degrading writings or speech, which inappropriately attacked certain groups or individuals. Much of this hate material was encountered on social media, such as Facebook, Twitter and YouTube (Hawdon et al. 2017).
- The percentage of survey respondents who reported having been *personally targeted* was much lower, at around **11 per cent**.[11] The percentage was highest in the US (16 per cent), followed by the UK (12 per cent), Finland (10 per cent) and Germany (4 per cent).
- Similarly, rates of sending hate material were low in the sample. Those in the US were most likely to admit to this act (**4.1 per cent**), followed by respondents in Finland (**4.0 per cent**), the UK (**3.4 per cent**) and Germany (**0.9 per cent**).
- Young men living alone with a close connection to the online world were most likely to post hate material (Kaakinen et al. 2018a).

Direct observation studies of online racial and religious hostility have found broadly similar results.

- One of the very first studies to measure the phenomenon at source, found 1,878 English language tweets were posted in the aftermath of the Woolwich terror attack in 2013 that constituted religiously (anti-Muslim) hostility. This amounted to just under **1 per cent** of all posts collected in the study period (Burnap and Williams 2015, Williams and Burnap 2016).
- In a study of geolocated tweets sent in London in an eight-month period between 2013 and 2014 (21.7 million), Williams et al. (2020) classified **1.4 per cent** as either anti-black or anti-Muslim.
- Research has also examined the online social reaction to the Brexit vote by monitoring the world-wide and UK-only posting of anti-Muslim content on Twitter, finding a similar percentage of hate speech (**2 per cent** world-wide, **1 per cent** UK) compared to total communications about the topic (Demos 2016a,b, 2017).
- Ozalp et al. (2020) studied online antagonistic content related to Jewish identity posted on Twitter between October 2015 and October 2016 using some UK specific keywords, finding **0.7 per cent** of the posts collected constituted antisemitic discourse.

[11] The sample did not specifically target those with protected characteristics.

- Overall, these results and others estimate that approximately **0.5-2 per cent** of Twitter based communications contain text that targets race or religion with a degree of hostility.

Correlates of Online Hate Speech

It is clear from existing research that 'trigger events' are the macro-off-platform correlates of online hate speech (see Conceptual Frameworks). However, research has found micro-on-platform correlates of online hate speech, based on an inspection of user, social, content and temporal factors.

Summary of key findings:

- Williams and Burnap (2016) found that in the aftermath of the Woolwich terror attack, of the various online users posting a reaction, those associated with the Far-Right were **three times more likely** to produce anti-Muslim posts compared to general Twitter users.
- Content features also emerged as significant correlates. The odds for including hashtags were higher for anti-Muslim tweets, while they were lower for URLs. The authors note this may suggest those wishing to promote hateful content online via contagion use hashtags to enhance the discoverability of their posts. Conversely, URLs are possibly less common in hateful tweets given linked content (most often a popular media source) is unlikely to corroborate hateful attitudes and biased speculative rumours.
- Temporal factors were also significant correlates. Hostile posts were more likely to be sent on the morning commute and in the evening.
- There was a positive association found between the publication of conventional press headlines about the Woolwich attack and the posting of anti-Muslim tweets. The odds of the production of extreme anti-Muslim posts increased by a magnitude of **1.3** for every 100 additional news headlines produced.
- This link between the frequency and number of newspaper headlines related to a terror attack and hate crime/speech has been replicated. Ivandic et al. (2019) found Islamic extremist terror attacks attract around **375 per cent** more media coverage than those based on other motivations, leading to a situation where the public have an inflated sense of threat from these types of attack. The researchers showed frequency in press reporting had a strong association with hate crimes on the street and online. Other research further corroborates this pattern (see Schmuck et al. 2020, Menshikova and Frank van Tubergen 2022).
- In a study of 35 UK-active Far-Right actors and their posts, Sprejer et al. (2022) found that the number of account followers best explained engagement, but the nature of those followers did not. Interacting with the audience by explicitly requesting retweets was also highly correlated with engagements. The authors also showed the type of tweet the actor produced had a large impact on engagement. Quoted posts and replies were associated with far less engagement than original posts. Actors who produced more toxic tweets on average received more engagement. The inclusion of media also had an impact, where posts with a video or image receive most engagement. Like Williams and

Burnap, Sprejer et al. also showed the number of hashtags had a significant but small effect.

- Williams and Burnap (2016) also found that tweets containing hate terms were significantly less likely to be retweeted and to survive on Twitter (duration), compared to tweets not containing such content. Posts sent by users associated with the Far-Right were the least likely to be retweeted. Conversely, tweets containing positive words and phrases (e.g. 'warm wishes to the family of Lee Rigby', 'brave family', 'respect for armed forces', etc.) as opposed to negative words or phrases were **38 per cent** more likely to be retweeted, and posts from news sources and police were **4.3** and **5.7 times** more likely to be retweeted compared to the average user.

Empirical Association Between Online and Offline Hate

Research on the link between online and offline phenomena falls into two categories: studies that hypothesise social media is a 'mirror' of offline behaviour and studies that hypothesise social media is a 'motor' for offline behaviour. The first position sees social media users as *sensors* of offline phenomena. For example, sensors can detect neighbourhood disorder via *signs* (e.g. graffiti, vandalism, litter, community tensions) and comment on these observations on social media (e.g. to raise community awareness, to report to local the council and/or police via their social media accounts). Users can publish information about local disorder in four ways: as victims; as first-hand witnesses; as second-hand observers (e.g. via media reports or the spread of rumour); and as perpetrators (Williams et al. 2020). This perspective assumes online behaviours have minimal influence on offline behaviour, and that what is posted online is merely a reflection of offline phenomena.

The alternative position postulates that social media can act as a 'motor' driving offline behaviours. In the case of polarisation, for example, opposing online communities emerge due to escalating reactions to divisive online debates fuelled by engagement algorithms that push content containing partisan views. Some users begin to imagine offline communities are as polarised as online communities, and building frustrations result in a form of expression in physical spaces (Sunstein 2017, Hassan et al. 2022). Investigations into recent Far-Right terror attacks lend support to this position.

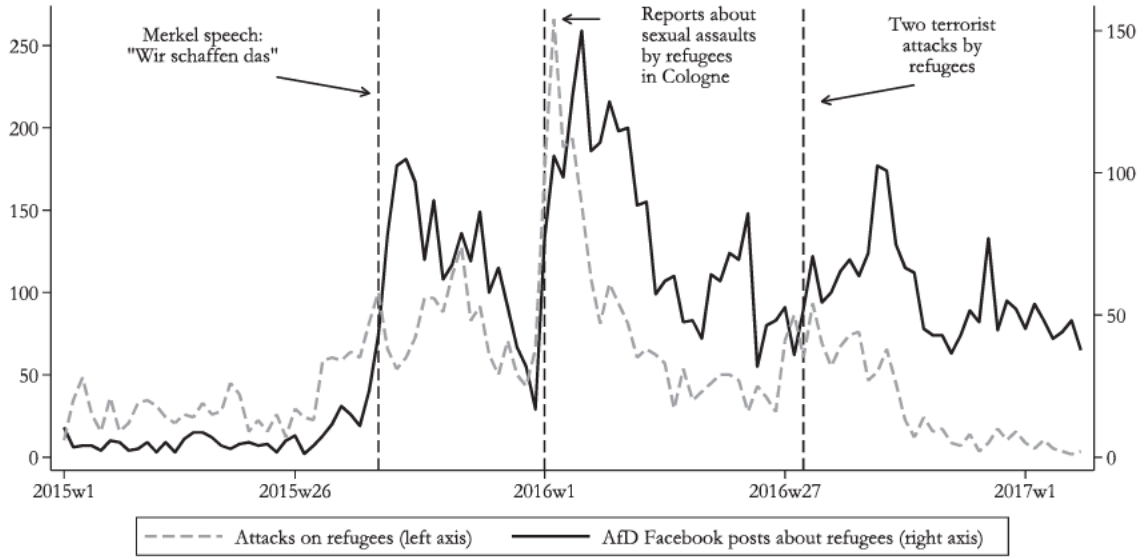
Of course, the separation of the 'mirror' and 'motor' positions inaccurately reflects reality, and social media acts as both, in a complex iterative fashion. Research is now generating empirical evidence on the link between online and offline hateful behaviours.

Summary of key findings:

- Müller and Schwarz (2021) found anti-refugee posts on the Far-Right Alternative für Deutschland (AfD) Facebook page triggered offline violent crime (especially assault) against immigrants in German municipalities (Figure 2). A one standard deviation higher Far-Right social media usage was associated with a **10 per cent** higher probability of an antirefugee incident on the streets. Facebook and internet outages saw the effect decrease by **18**

and 53 per cent respectively. The authors argued that social media is not directly causal of hate crimes but may motivate collective offline action.

Attacks on Refugees in Germany and AfD Facebook Posts About Refugees (Müller and Schwarz 2021)

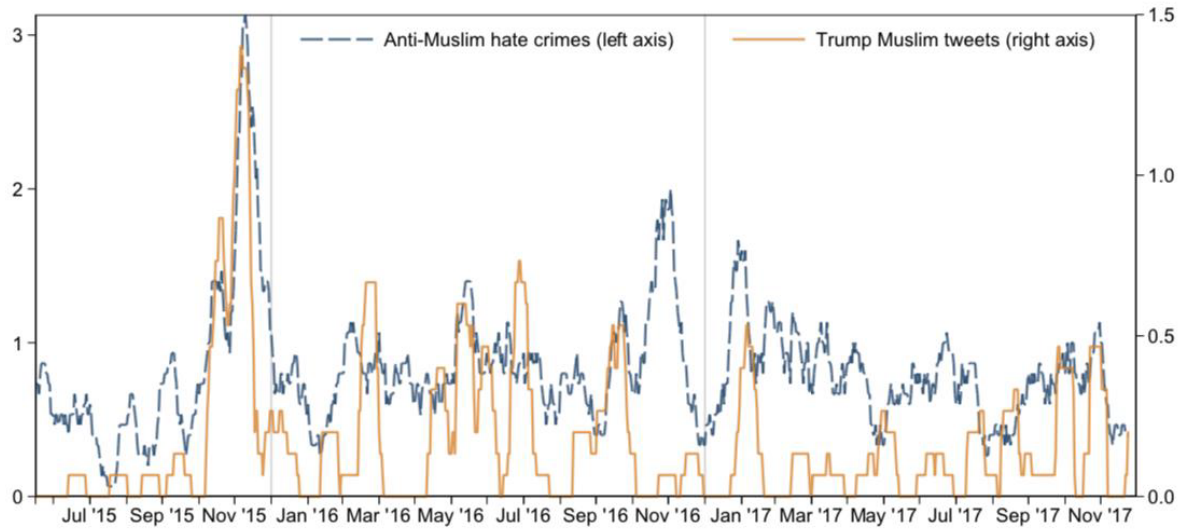


- The same authors also found a strong statistical association between President Donald Trump's tweets about Muslims and offline anti-Muslim hate crime in US counties (Müller and Schwarz 2020).[12] Trump's tweets about Muslims were almost **twice as likely** to be retweeted by his followers than his tweets on any other subject. His divisive anti-Muslim rhetoric also caused a **58 per cent** spike in other Twitter users posting with the hashtags #BanIslam and #StopIslam. No increase was found in anti-Muslim posts before Trump's tweets.
- Figure 3 shows Trump's tweets containing reference to Muslims (solid line) and anti-Muslim police-recorded hate crimes in the US (dashed line) between 2015 (wk 26) and 2016 (wk 50). The variation in the frequency of Trump's Muslim-related tweets and hate crimes offline is remarkably similar. This correlation could reflect that Trump reacted to US-wide anti-Muslim hate crimes driven by factors like terrorist attacks. Equally, it could be that Trump's anti-Muslim tweets encouraged those with existing prejudices to commit offline hate crimes.
- To test which of these explanations was most likely, Müller and Schwarz controlled for a wide range of other potential contributory factors, including (at the county level) population growth, age, ethnic composition, number of hate groups, qualifications, poverty rate, unemployment rate, local income inequality, share of uninsured individuals, household income, vote share of the Republican party, salience of Muslim-related topics based on Google searches, viewership

[12] Trump is not the only politician to use hate speech that is linked to hate crimes on the streets. Research across 163 countries covering the period 2000 to 2017 found that the use of hate speech in speeches by mainstream politicians was linked to higher rates of domestic political violence. J. A. Piazza, 'Politician Hate Speech and Domestic Terrorism', International Interactions 46 (2020), 431–53.

share of Fox News, cable TV spending, prime-time TV viewership and the number of mentions of Muslims on main US TV networks (Fox News, CNN and MSNBC). Their results confirmed that Trump's anti-Muslim tweets pre-dated the hate crimes, but only for the time period after the start of his presidential campaign.

Anti-Muslim Hate Crimes in USA and Trump's Muslim-Related Tweets (Müller and Schwarz 2020)



- A positive, but weaker effect was also found in relation to Trump's anti-Latinx tweets and offline anti-Latinx hate crimes during the same period. The pattern was mainly driven by cases of assault and vandalism. This gives reassurance that the finding is not a function of Trump's tweets increasing reporting. If that were the case, more reports on hate crimes of lower significance, such as minor public order offences would be observed.[13] In addition, the US National Crime Victimisation Survey did not show an increase in reporting from hate crime victims during the period in question.
- In the UK, Williams et al. (2020) found a similar link between online and offline hate in London. Anti-Muslim and anti-black hate speech posted on Twitter by ordinary users correlated with racially and religiously aggravated hate crimes on the streets. An increase of 1,000 hateful tweets corresponded to a **4 per cent** increase in racially or religiously aggravated harassment within a given month within a given Lower Layer Super Output Area (LSOA) in London. An interaction term between hate speech count and BAME population was significant, showing in an LSOA with a 70 per cent BAME population with 300 hate tweets posted a month, the incidence rate of racially and religiously aggravated violence was predicted to be between **1.75 and 2**. The association may point to rising levels of collective racial and religious tension that first erupts online, and then migrates onto the streets if not addressed.

[13] To further rule out reverse causality – i.e. the possibility that anti-Muslim hate crimes resulted in Trump tweeting about Muslims – the researchers turned their focus to when Trump tweeted. He was found to tweet about Muslims significantly more on golfing trips, when away from Washington and hence policy issues, and when in the presence of his social media manager, Dan Scavino, who is known to have suggested divisive tweet topics to Trump. Trump's planned golfing trips are clearly independent of spikes in hate crimes, indicating it is unlikely that his anti-Muslim tweets were driven by a possible knowledge of a rise in hate crimes.

- The authors state those doing the hateful tweeting and those committing the offline hate crimes may be different people, but it is most likely that there is a mix of those who did both and those who did only one of the activities. The authors conclude that when the number of hate tweets reaches a certain level in a London area where there are a significant number of BAME residents, hate crimes on the streets are more likely. This may be a useful finding for policy and practice as such estimations can alert social media users, Twitter and the police to implement counter-measures in an attempt to prevent offline hate crimes.

Conceptual Frameworks

Trigger events

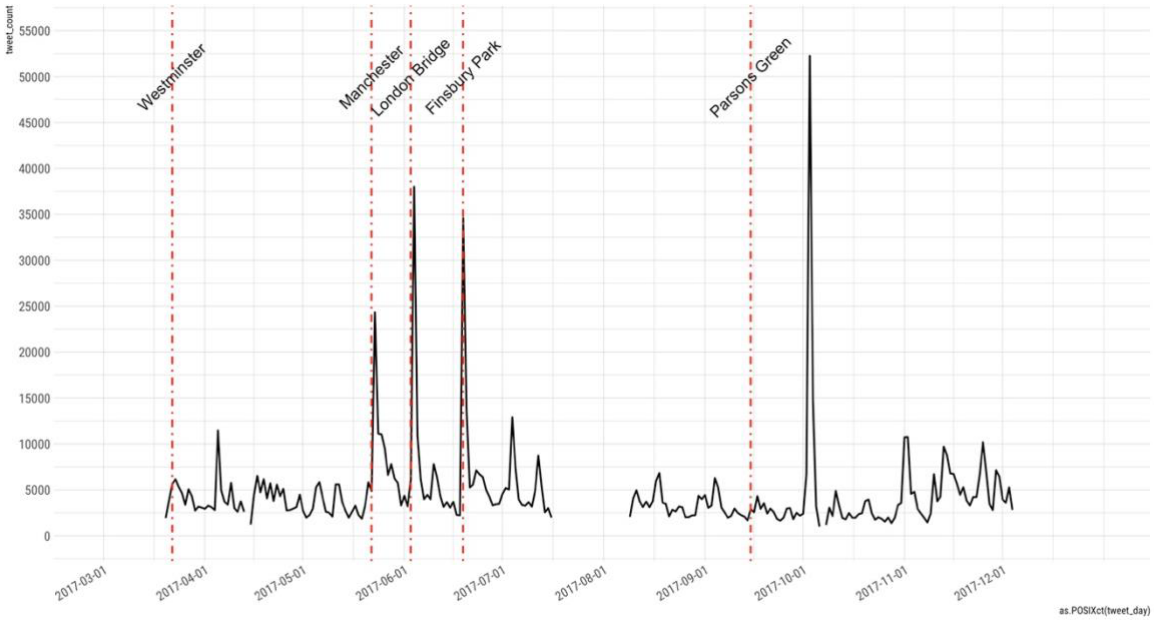
Research indicates that the expression of prejudiced opinions and violent acts of hate are linked to temporal variation in perception of outgroup threat. Trigger events, such as terror attacks, controversial political votes, and high-profile court cases, are known to impact on the perception of threat from an outgroup.

Summary of key findings:

- Legewie (2013) established a significant association between anti-immigrant sentiment and the Bali terrorist bombing. The Bali attack was the cause of a significant worsening in attitudes towards immigrants in Portugal, Poland and Finland. The strength of the effect of the attack was enhanced if a person lived in an area with high unemployment – both Poland and Portugal showed the highest increase in unemployment in 2001–2. The effect was also stronger on people who did not have immigrants as friends or co-workers but who lived in areas with high immigrant numbers. These findings were replicated in relation to the Islamist terrorist bombing in Madrid in 2004. The percentage of the Spanish population that thought immigration was one of the most important issues the country was facing increased from **8 to 21 per cent** immediately after the attack, with the effect being strongest in areas with high unemployment.
- King and Sutton (2014) found an association between terrorism and hate crime incidents in the United States. Following the 9/11 terrorist attack, law enforcement agencies recorded 481 hate crimes with a specific anti-Islamic motive, with **58 per cent** of these occurring within two weeks of the attack (4 per cent of the at-risk period of 12 months).
- In the United Kingdom, Hanes and Machin (2014) found significant increases in hate crimes reported to the police in London following 9/11 and 7/7. A sharp de-escalation was evident following the spike in hate crimes following the trigger events, indicating that event-specific motivated hate has a ‘half-life’. The authors conclude hate crimes cluster in time and tend to increase, sometimes dramatically, in the aftermath of antecedent ‘trigger’ events, such as terrorist acts. Edwards and Rushin (2019) later replicated these results finding a sharp increase in anti-Muslim hate crimes in the United States during the presidential campaign and subsequent election of Donald Trump.

- Williams et al. (2022) studied the role of the Brexit vote on hate crime rates across the four nations of the United Kingdom. When controlling for multiple factors (including variation in reporting, unemployment rate, average income, educational attainment, health deprivation, general crime rate, barriers to housing and services, quality of living, rate of migrant inflow, and Leave vote share), the referendum as a trigger event emerged as a powerful explanatory factor. The month after the vote saw **1,100 more** hate crimes (**29 per cent** greater) than would have been expected in the same period in the absence of the vote. Of the other factors that may have had an impact, only the Leave vote share emerged as a significant predictor of hate crimes. The higher the Leave vote in an area, the greater the increase in hate crimes after the vote.

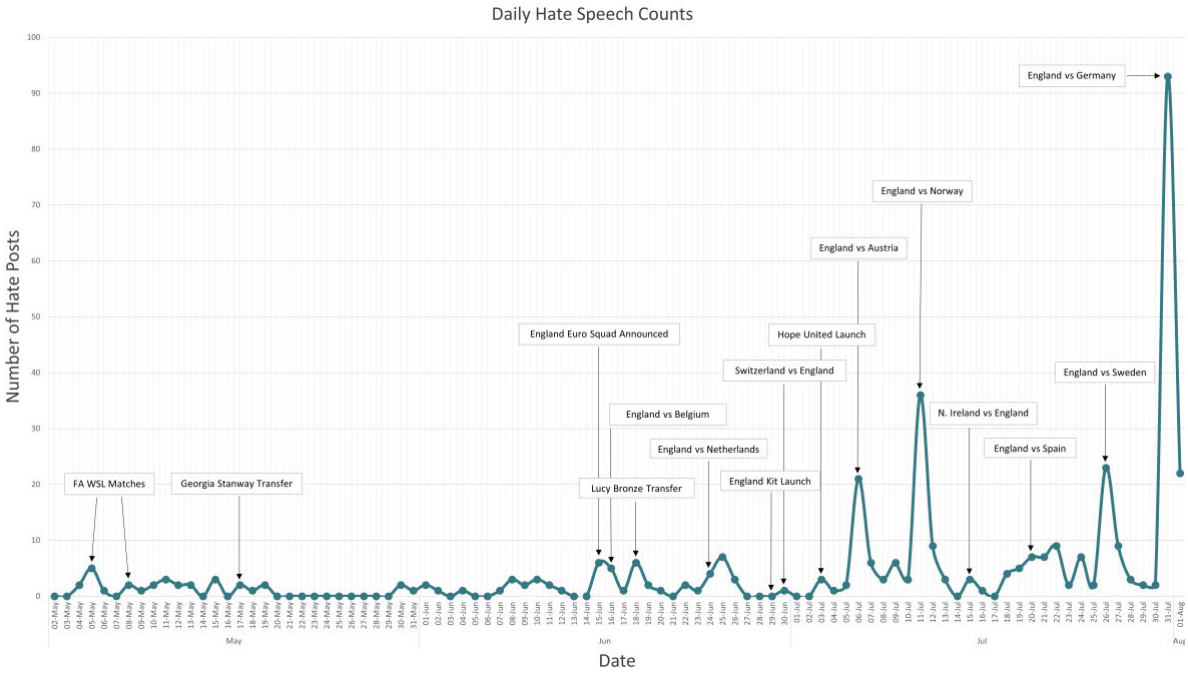
Online Anti-Muslim Hate Speech posted on Twitter in 2017 (Williams et al. 2020)

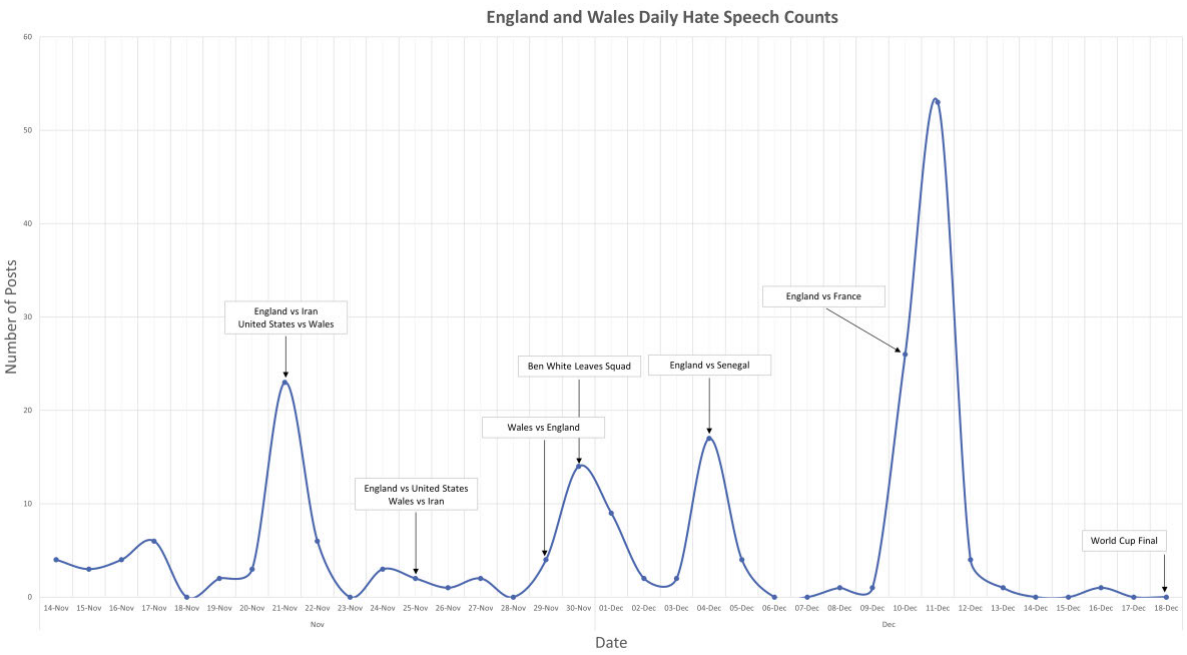


- Research has also found offline trigger events are associated with variation in the production of online hate speech. Williams et al. (2016, 2020) were the first find an association between terror attacks and online hate speech. They found the Woolwich, Manchester, London Bridge and Finsbury Park terror attacks were followed by and a rapid increase, and subsequent sharp decline, in online anti-Muslim hate speech (see Figure 4). These results have since been widely replicated across the globe (see Kaakinen et al. 2018b, Innes et al. 2018, Roberts et al. 2018, Fischer-Preßler et al. 2019, Bérubé et al. 2020, Scharwächter and Müller 2020, Álvarez-Benjumea and Winter 2020, Gallacher and Heerdink 2021, and Vidgen et al. 2022).
- Controversial statements by political figures have also emerged as triggers, with Ozalp et al. (2020) finding a significant association between the Labour Party antisemitism claims and an increase in antisemitic hate speech on Twitter (see also Scrivens et al. 2021).

- Sporting events have also been identified as triggers for online hate. Williams (2021) found an association between the EURO 2020 final between Italy and England and racist abuse targeted at three black England players. Similarly, Cullen and Williams (2022, 2023) found an increase in online hate posts around England games in the Women's EURO 2022 and England and Wales games in the World Cup 2022 (See Figure 5).

Online Hate Speech Sent to Players from Home Nations During the Women's EURO 2022 (top) and World Cup 2022 (bottom)





Counter-Measures

Automated Moderation, Warning Prompts and Other Tools

Platform moderation can be delineated between ex ante and ex post – moderation takes place before or after posting (Grimmelmann 2015). The most common strategy for moderating hate-related content is to automatically classify and remove it after posting. This ex post approach relies on the use of machine learning to identify and prioritise content sometimes for later human moderation (Chandrasekharan et al. 2017, Wulczyn et al. 2017, Yin et al. 2009). This successfully removes billions of posts a year across platforms, and when accompanied with transparency, can reduce future hate posting behaviour.

Summary of key findings:

- Studies show that providing transparency in moderation decisions increases future rule compliance. Tyler et al. (2019) demonstrate that posters who had content removed were less likely to post content that broke community guidelines in the future when provided the reason for the sanction. Similarly, Jhaver et al. (2019) found that when provided an explanation for a removal, users on Reddit were less likely to post comments that broke platform rules in the future.
- As an alternative, some large platforms have experimented with the use of warning prompts delivered on the user's device just before the point of sending, with the aim of nudging them away from posting content that may contravene hateful conduct policies. Katsaros et al. (2022) found the use of such a nudge on Twitter resulted in **6 per cent** fewer offensive tweets being posted than in the control group with no nudge, with **9 per cent** of nudged posts cancelled and **22 per cent** revised (37 per cent of these were to a less offensive alternative) before sending.[14] Decreases were also recorded in the number of nudged users creating offensive posts in the future **(<20 per cent** compared to the control group) and the number of offensive replies to nudged posts. The decrease in offensive replies sent by nudged users was not followed by a reduction in total replies sent. This suggests ex ante interventions that decrease a user's offensive contributions do not come at the cost of reducing healthy free-speech on the platform. These ex ante approaches have been rolled out across many platforms, including Instagram,[15] YouTube,[16] Pintrest,[17] and TikTok.[18]
- A range of platforms have also introduced tools that reduce the ability to directly target individuals with hate speech. For example, Twitter introduced a new feature to turn off @mentions on posts via the Unmention tool, YouTube introduced the ability to tun off comments on video posts, and TikTok allows

[14] The effectiveness of the intervention increased amongst Portuguese users in Brazil.
[15] https://about.fb.com/news/2019/12/our-progress-on-leading- the-fight-against-online-bullying/
[16] https://blog.youtube/news-and-events/new-tools-to-shape-conversations-in/
[17] https://newsroom.pinterest.com/en/creatorcode
[18] https://newsroom.tiktok.com/en-us/new-ways-to-foster-kindness-and-safety-on-tiktok

creators to filter comments by excluding certain keywords, such as slurs and epithets.

Content Referral

Since 2015, platforms have built specific teams and resources for addressing hate speech and violent extremism. Experts have been recruited from law enforcement, law, academia, and government to inform the development of content moderation policies. Large moderations teams have been established in Meta, Twitter, Google, TikTok and many others. For example, Meta is reported to employ (directly and through third parties) ~15,000 content moderators across multiple regional teams based in over 20 locations. All the large platforms have policy collaborating with technical teams to develop AI approaches to enhance content moderation practices. AI flags and removes the majority of content that violates platform conduct policies.

Beyond AI, each platform has its own mechanism for the referral of content by ordinary users. Increasingly, platforms are employing in-situ methods where users are asked if they wish to report content as potentially harmful in some way. The speed of response following a referral from a general user varies significantly between platforms. For example, research shows Meta generally provides a removal decision within 24–48 hours. Decision to remove content also varies by platform, and research suggests a **~50 per cent** Meta deletion rate (Carlson and Rousselle 2020). Legal demands, both civil and criminal, can also be made.

Twitter's July to Dec 2021 transparency report shows that there was a total of 16 legal demands made from UK organisations, with a compliance rate of **37.5 per cent**. To put this in context, in the same period Twitter received 47,527 legal demands globally, with a **51 per cent** compliance rate. Five countries were responsible for **97 per cent** of the total global volume of legal demands: Japan (50 per cent), Russia (18 per cent), South Korea (12 per cent), Turkey (9 per cent), and India (8 per cent). Facebook and Google do not provide such detail in their transparency reports.

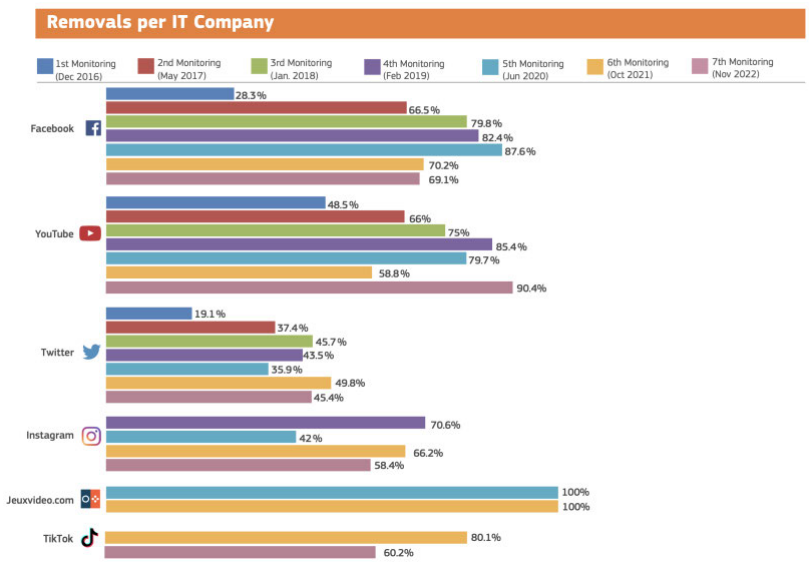
Many social media companies have a network of 'trusted flaggers', organisations that have a mandate to report content that may be considered illegal hate speech under local European laws, which have formed formal partnerships with platforms. These networks were set-up as part of the continuing efforts under the EU Code of Conduct on Countering Hate Speech Online. Initially in 2016, Facebook, Microsoft, Twitter and YouTube signed up to a code of conduct, with Instagram, Google+, Snapchat and Dailymotion joining later in 2018. By signing they all agreed on rules banning hateful conduct and to introduce mechanisms, including dedicated teams, for the review and possible removal of illegal content within 24 hours.[19]

[19] The organisations only notify the platforms about content deemed to be illegal hate speech under national laws.

Figure 6 shows the compliance rates from the last 7 evaluations of the scheme.[20] The latest results from the monitoring exercise between 28th March to 13th May 2022 show a fall in performance:

- Posts *reviewed* within 24 hours dropped compared to the last 2 years, from **90 per cent** in 2020, to **81 per cent** in 2021, and **64 per cent** in 2022.
- The *removal rate* was **64 per cent** in 2022, lower than the peak in 2020 of **71 per cent**.
- Compared to the year before, all but YouTube and Jeuxvideo saw their removal rate drop in 2022.
- In 2016, when monitoring first began, only **40 per cent** of participating platforms *reviewed* the majority of hate posts flagged to them within 24 hours, and **28 per cent** of these posts were *removed*.
- In 2022, Xenophobia (including anti-migrant hatred) and homophobia were the most commonly reported grounds for hate speech (**16.3 per cent** and **15.5 per cent** respectively).
- Across platforms, **70 per cent** of content calling for murder or violence against specific groups was removed, while content using slurs, defamatory terms or pictures to describe certain groups was removed in **59 per cent** of cases.
- Compared to the year before, platform **feedback to users** improved in 2022, with TikTok (**75 per cent** of notifications addressed, compared to **29 per cent** in 2021) and Instagram (**73 per cent**, compared to **42 per cent** in 2021 and **62 per cent** in 2020) improving markedly.
- The average removal rate for the UK was **43 per cent** in 2022, the same as in 2021. This compares to **77 per cent** in Germany in 2022, and **95 per cent** in 2021.

Evaluations of the EU Code of Conduct on Countering Hate Speech Online



[20] The 7th exercise was carried out for a period of approximately 6 weeks, from 28 March to 13 May 2022. The figures do not intend to be statistically representative of the prevalence and types of illegal hate speech in absolute terms, and are based on the total number of notifications sent by the organisations in the trusted flagger network.

Country comparisons are important as they may indicate how existing social media laws are working to change the compliance of platforms. The Netzwerkdurchsetzungsgesetz (NetzDG) law in Germany, that came into force in October 2018, imposes fines of up to 50 million euros on social media companies if they fail to remove illegal hate speech flagged by a trusted third party in the country. Durán et al. (2022) examined the effect of the NetzDG on the amount of hateful posts online. They found that its introduction was associated with a statistically significant reduction in hate posts by Far-Right social media users and a reduction in anti-refugee hate crimes in towns with more Far-Right Facebook users.

Online-Counter-Speech

Counter-speech seeks to undermine hateful criminal speech or legal but harmful speech via a direct or general response. The average social media user can use counter-speech to influence online discourse, sometimes resulting in a desistance effect on the speaker and/or 'inoculating' the audience against hate speech so they are less easily influenced by it (Williams and Mishcon de Reya 2019).

Combating hate speech with counter-speech has some advantages over platform and law enforcement sanctions:

i) It can be rapid
ii) It can be adapted to the situation
iii) It can be initiated by any internet user

HateLab is currently testing the effectiveness of different types of counter-speech sent to those posting hate speech on Twitter. Six forms of counter-speech are being considered:

- Attribution of prejudice and moral suasion
  e.g. "Shame on you for spreading sexist tropes like that! Imagine if someone said that about your daughter."

- Claims making and appeals to reason
  e.g. "This has nothing to do with immigration! Take a look at these statistics."

- Request for information and evidence
  e.g. "How does this have anything to do with religion?? Do you have any proof?"

- Jokes/comedy and reintegrative shaming
  e.g. Oh no, this trans woman sounds REALLY scary. I'm going to join a TERF group immediately - you guys. LOL!

- Mimicry and sarcasm highlighting issues with logic and consistency
  e.g. Hate speech: "I'm officially scared of butch lesbians. #NotHomophobic #JustScared"

Mimicry: “I’m officially scared of bigoted men. #StereotypingMuch? #JustStupid”

- Reductio Ad Absurdum (an argument pushed to its absurd extremes to identify its inherent problem)
  e.g. “I guess what you’re saying is you would feel more comfortable if women didn’t exist? Fascinating. Do you want the human race to go extinct?”

A body of evidence is emerging on the effectiveness of these various interventions.

Summary of key findings:

- To test if online counter-speech was useful in positively engaging with online right-wing extremists, researchers at the Institute for Strategic Dialogue measured initial response rates, sustained engagement and indicators that the candidate was questioning their online behaviour. Results showed that **16 per cent** of candidates responded to the initial outreach contact, and that an argumentative tone produced the greatest success, followed by casual, meditative and reflexive tones. A scholarly tone yielded no response. Of those who responded, **64 per cent** engaged in sustained conversation, and **6 per cent** subsequently questioned their online behaviour (ISD 2018).
- HateLab (2019) conducted a mixed methods study of responses to online hate posts in order to understand interactional dynamics between hate-speech and counter-speech producers. They found counter-speech was effective in stemming the length of hateful threads on Twitter, and that the effect was strengthened when more counter-speakers entered into the interaction.
- In one of the first counter-speech quasi-experiments, Álvarez-Benjumea and Winter (2018) compared two types of interventions: counter-speaking (informal verbal sanctions) and censoring (deleting hateful content), finding users were significantly less likely to engage in hate speech when prior hate content had been censored (deleted).
- Following this study, Munger (2019) conducted an experiment where bots were used to sanction Twitter users who posted racist content. The identities of the bots were varied between in-group (white man) and out-group (black man). The results showed those sent counter-speech by an in-group bot (white male) significantly reduced their use of a racist slurs for a two-month period. Munger (2021) conducted a follow-up experiment where counter-speech posts were varied by their moral content. Posts using moral suasion were more effective at stemming the production of partisan incivility on Twitter compared to posts that did not contain moral content.
- Siegel and Badaan (2020) examined the effect of counter-speech narratives in the Arab Twittersphere via a nationally representative survey experiment. They found that elite-endorsed messages that primed common religious identity were most effective in reducing hate speech, suggesting elites play an important role in alerting individuals to social norms of acceptable behaviour.

- Kunst et al. (2021) examined the role of solidarity citizenship norms online, showing users who supported these norms had a greater tendency to flag hate comments and to engage in counter-speech.
- Bilewicz et al. (2021) conducted an experiment on Reddit with three counter-speech interventions delivered by bots: induction of a descriptive norm, induction of a prescriptive norm, and empathy induction. All three interventions proved effective in reducing online aggression when compared with the control condition (no counter-speech).
- A study by Hangartner et al. (2021) randomly assigned posters of racist hate speech to one of three counter-speech strategies (empathy, warning of consequences, and humour) and a control (no counter-speech). Their results showed that while there was no effect for strategies using humour or warning of consequences, empathy-based counter-speech messages increased the retrospective deletion and reduced the future posting of racist hate speech over a 4-week follow-up period. Empathy based counter-speech humanised the victim by highlighting that they were harmed by the hateful post.
- Durán (2021) found that randomly reporting posts on Twitter for violating the rules against hateful conduct increased the likelihood that they were removed. However, reporting did not change the likelihood of post authors reposting hate, but it did increase the counter-speech activity of those attacked by the posts. The study suggests that content moderation does not necessarily moderate the behaviour of hateful posters.
- Garland et al. (2022) conducted a large-scale longitudinal study of the dynamics of hate and counter-speech in German political discussions online. They found that organised counter-speech appears to contribute to a more balanced public discourse. After the emergence of Reconquista Internet (RI), an organisation who actively resist hate discourse, the relative frequency of counter-speech increased while hate speech decreased. Counter-speech became more effective in steering conversations when organised through RI, primarily by providing more support to counter-speakers and by steering information flows towards neutral discourse. While the tactics of RI were met with a backlash initially, the relative frequency of hate speech stabilised to a lower baseline.

Counter-Speech Generated by AI

- A study by Costello et al. (2024) investigated whether dialogs with a generative artificial intelligence (AI) interface could convince people to abandon their conspiratorial beliefs. Human participants described a conspiracy theory that they subscribed to, and the AI then engaged in persuasive arguments with them that refuted their beliefs with evidence. The AI chatbot's ability to sustain tailored counterarguments and personalised in-depth conversations reduced their beliefs in conspiracies for months, challenging research suggesting that such beliefs are impervious to change. The treatment reduced participants' belief in their chosen conspiracy theory by 20% on average. This effect persisted undiminished for at least 2 months; was consistently observed across a wide range of conspiracy theories, and occurred even for participants whose

conspiracy beliefs were deeply entrenched and important to their identities. Notably, the AI did not reduce belief in true conspiracies. The debunking also spilled over to reduce beliefs in unrelated conspiracies, indicating a general decrease in conspiratorial worldview, and increased intentions to rebut other conspiracy believers.

- A study by researchers in the University of Zurich (2025) found that AI generated counter-narratives surpassed human performance substantially, achieving persuasive rates between three and six times higher than the human baseline. The study demonstrates that AI can be highly persuasive in real-world contexts. While persuasive capabilities can be leveraged to promote socially desirable outcomes, their effectiveness also opens the door to misuse, potentially enabling malicious actors to sway public opinion or orchestrate election interference campaigns.

Initial research suggests the wide adoption of counter-speech would see a reduction in hateful communications on large platforms. Those most susceptible to the stemming effects of counter-speech are users who engage in hate speech only occasionally (for example, around 'trigger' events – defensive, retaliatory and thrill-seeking posters – see Williams 2021). General counter-speech is unlikely to stem the production of hate speech in hardened 'mission hate' posters. Counter-speech is also unlikely to be effective on some bots and fake accounts, given their control is either fully or partially automated by computer code (Williams and Mishcon de Reya 2019). No evidence exists to suggest counter-speech from general social media users is more or less effective than counter-speech used by government officials or police. Furthermore, emerging research indicates that AI generated counter-speech is as effective, if not more effective, than human generated counter speech at challenging existing beliefs and world-views.

Appraisal of Counter-Measures

- Automated content moderation is the most common counter-measure used by platforms, and results in millions of deletions a day facilitated by machine learning algorithms that have been trained to identify content that breaches platform policies. This counter-measure is most likely to generate large amounts of false negatives (missing some legal but harmful content) and false positives (removal of content that is not harmful). Despite these limitations, automated moderation is an important counter-measure given its ability to deal with the scale and speed of posts generated by large platforms.
- Automated prompting (nudges) show promise in diverting users away from behaviours that violate platform policies. More experimental research is needed, in partnership with platforms, to fully evaluate this counter-measure.
- Content perceived as problematic is frequently referred to platforms by organisations (such as NGOs, police) and users, which is reviewed by machine learning algorithms, and if necessary human moderators. Not all referrals result in action (such as suspension or deletion), but the processes now put in place by larger platforms mean that content missed by automated counter-measures can be dealt with by referral. Larger and more diverse moderation teams (in terms of

language) will improve the capability of platforms to deal with referrals more rapidly and with higher accuracy.

- Counter-speech that can be used by organisations and everyday users, has been found to be effective in stemming the production of hateful posts. The limited evidence suggests that counter-speech using moral suasion and empathy induction, delivered by a member of an ingroup or a respected (elite) figure, is most likely to succeed in changing the behaviour of hateful posters.
- Those most susceptible to the stemming effects of counter-speech are users who engage in hate speech only occasionally (for example, around 'trigger' events – defensive, retaliatory and thrill-seeking hate posters).
- Emerging evidence suggests counter-speech generated by AI is effective in challenging hate speech and conspiracy theories, resulting in opinion and behaviour change.
- General counter-speech is unlikely to stem the production of hate speech in hardened 'mission hate' posters. Counter-speech is also unlikely to be effective on some bots and fake accounts.
- More robust research is needed to fully assess the effectiveness of all of these counter-measures, and to determine which combination is optimal.

Application Areas

Monitoring Online Hate Speech

There continues a policy and practice need to improve the intelligence on hate crime, and in particular to better understand the role community tensions and events play in patterns of perpetration. The HMICFRS (2018) inspection on police responses to hate crimes evidenced that forces remain largely ill-prepared to handle the dramatic increases in racially and religiously aggravated offences (on and offline) following events like the United Kingdom European Union referendum vote in 2016 and the terror attacks in 2017.

Part of the issue is a significant reduction in Police Community Support Officers throughout England, and in particular London (Greig-Midlane (2014) indicates a circa 50 per cent reduction since 2010). Fewer officers in neighbourhoods gathering information and intelligence on community relations reduces the capacity of forces to pre-empt and mitigate spates of inter-group violence, harassment and criminal damage.

Technology has been heralded as part of the solution by transforming analogue police practices into a set of complementary digital processes that are scalable and deliverable in near real-time (Chan and Bennett Moses, 2017; Williams *et al*. 2017a). Over the past decade, various technologies have been developed within corporate and academic research settings that facilitate the real-time and historic monitoring of open-source online communications for the purposes of promoting community cohesion and tracking extremist activity.

An early government response was to establish the National Online Hate Crime Hub in 2016, which has been supported by publicly funded technology from Cardiff University's HateLab. This technology has been deployed during several operations (e.g. 'Punish a Muslim Day', LGBT Pride Month 2022, online abuse of Members of Parliament), resulting in improved targeted police responses and significant resource savings.

Commercial Solutions

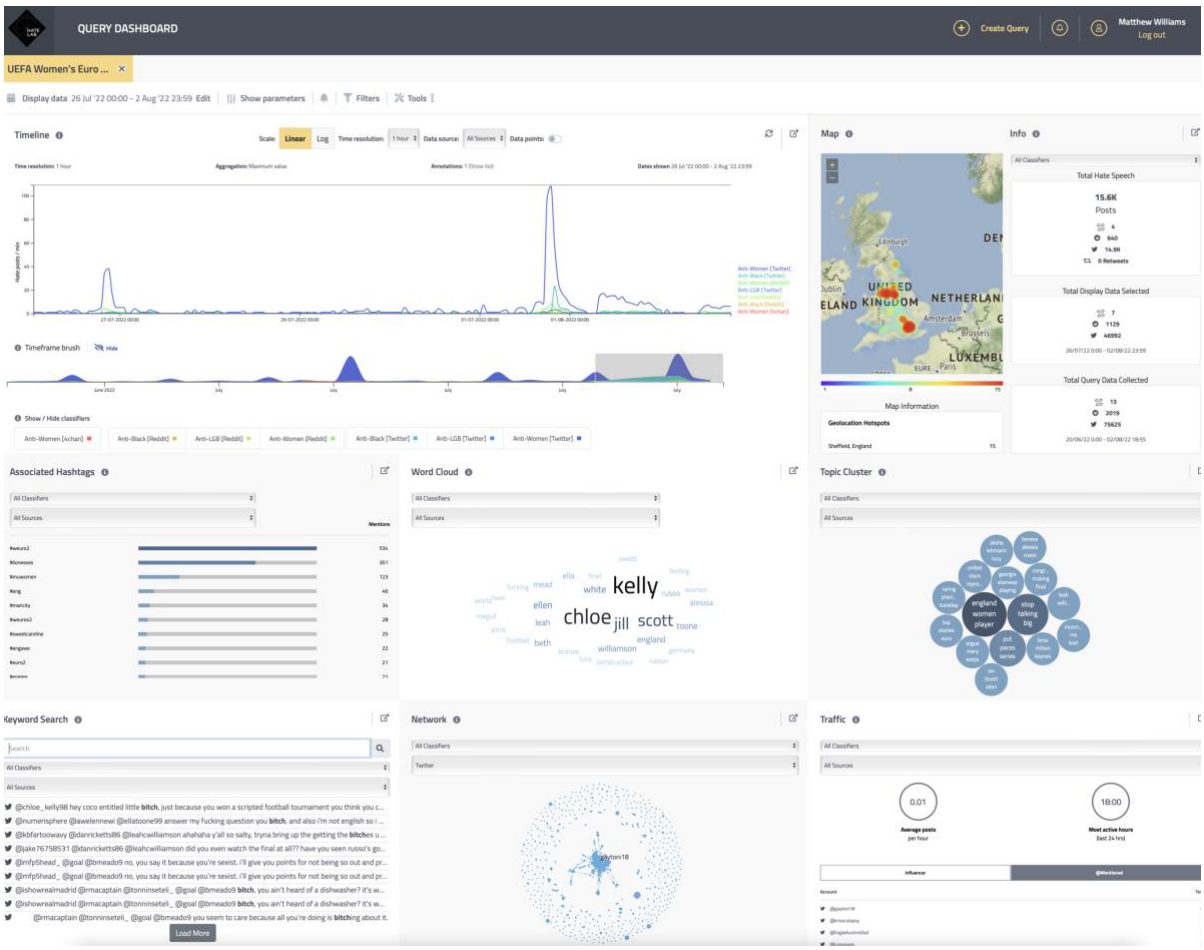
Over the past few years, the UK Safety Tech sector has grown. The 2022 DCMS UK Safety Tech Sector Analysis found the number of operating firms grew 17 per cent in the last year, to 117. System wide governance firms, those which provide technology for the automated identification of illegal content (such as CSEA material and terrorist content), numbered 14 firms, while platform level firms, those which support content moderation through identifying and flagging content to human moderators, numbered 29 firms.

Examples of trust and safety sector firms include Nisos, which helps platforms defend against and respond to advanced cyber attacks, disinformation, and abuse, Pattr, which safeguards users and brands from toxicity and spam in social conversations through the use of AI-driven tools and insights, ConteX.ai, which emulates human understanding in detecting toxicity across text, speech and images, and nisien.ai, which provides monitoring and mitigation AI products and services related to a wide variety of online harms that target people, companies, brands, products and campaigns.

Publicly Funded Solutions

Several publicly funded (UKRI) units are also active in the space. Cardiff University's HateLab is the longest standing research and practice unit that has studied the manifestation, spread and countering of online harms, including hate speech. Established by an ESRC-Google grant in 2013, HateLab's core objective is to develop and democratise technology amongst organisations that do not have the resources to routinely monitor and counter online hate speech and divisive disinformation. Its award-winning AI tech (see Figure 7) has been embedded within a range of policing, government and civil society organisations which have a remit to protect minorities, depolarise debates and strengthen democracies. HateLab's Dashboard, now part of it's spinout, Nisien.ai, has uplifted hate monitoring and mitigation capabilities in many organisations, resulting in successful counter-speech campaigns and significant resource and cost savings.

HateLab/nisien.ai's HERO Detect



Countering Online Hate Speech

Existing Initiatives

Several UK based organisations have pioneered counter-speech initiatives, including HateLab, the Institute for Strategic Dialogue and Demos. One notable success was spear-headed by Stop Hate UK. In 2017 they established and trained a team of counter-narrative speaker volunteers who worked in shifts to monitor and counter hate speech across platforms. The project was a success in terms of testing of a range of interventions while maintaining a focus upon the safety, security and well-being of counter-speakers. Anecdotal evidence showed the counter-narrative work was effective in addressing the online behaviour of some users posting hostile communications that did not breach the threshold of criminal legislation or social media hateful conduct policies. It allowed for these 'legal but harmful' posts that targeted protected characteristics to be addressed in the online 'town hall' that could be seen by significant numbers of 'bystanders' in addition to those targeted individually or collectively, promoting the idea of an 'online citizenship' and a shared, collective responsibility to call out hate.

While reported as a success, the first Stop Hate UK counter-narrative project was not fully evaluated. Planning of an evaluation framework should be designed into any future government supported project from inception to allow for the collection of data generated by volunteers and subjects to test, analyse and conclusively establish the effectiveness and impact of range of online interventions across platforms. From this work an online interventions best practice guide should be produced to inform practitioners, community groups and wider interested parties. Any 'intelligence' obtained as a by-product of any activity should be made available to the NPCC Hate Crime Gold Group to support a holistic response to hate crimes and incidents.

Nisien.ai is building a suite of Online Trust & Safety products using advanced AI, to help foster more constructive, factual and healthy online conversations. Nisien has received support from Welsh Government's SMART FIS to build its disruptive *HERO: Resolve* product, that automates the production of de-polarising counter-narratives with Generative AI. It also receives support from the AIRBUS Endeavr Wales Challenge programme to develop its *GenAI Detect* product, which identifies content created and modified by AI, including mis and disinformation. Both products will go to market in Q1 2026.

Conclusions

- Counter-measures are effective in stemming the production and spread of hateful speech online. Automated content moderation is the most common counter-measure used by social media Trust and Safety teams and is largely effective when fully deployed given its ability to deal with the scale and speed of posts generated by large platforms (Tyler et al. 2019, Jhaver et al. 2019). Content referral is also an effective counter-measure, but it is reactive and hence it takes more time for action to be taken, often meaning harm has already been inflicted upon the victim and/or the community (Carlson and Rousselle 2020). Content moderation and referral processes are largely controlled by platforms, meaning users and governments have limited influence. Emerging evidence suggests that counter-speech that uses moral suasion and empathy induction, delivered by multiple members of an ingroup, is most likely to succeed in changing the behaviour of hateful posters (Munger 2019, 2021, HateLab 2019, Siegel and Badaan 2020). As a community-based measure, counter-speech could have a significant impact on hate speech if well-orchestrated on a large scale. The automation of counter-speech using Generative AI has the potential to address online hate speech and resulting polarisation at scale (Costello et al. 2024).

Appendix: Use cases of HateLab/Nisien.ai HERO Detect

HERO Detect provides stakeholders with an essential service to identify and counter online hate speech. Access to real-time public communications from social media platforms allow analysts to monitor hate speech at an aggregate level using cutting-edge AI. To date, there have been significant delays in getting information to police and other emergency responders following 'trigger' events. HERO Detect allows key personnel to gain aggregate insights into online reactions to events, such as the aftermath of the Southport attack, in the so called 'golden hour'. The hour following a hate speech 'trigger' event represents a critical period for who we call 'online first responders'. Analysis shows that social media posts from the press, hate crime charities and police gain significant traction in the immediate aftermath of 'trigger' incidents. As a result, they have an opportunity to engage in counter-speech messaging, such as dispelling rumours and challenging stereotypes, to support victims and stop hate speech from spreading.

This appendix contains testimonials from clients of HateLab and its spinout, nisien.ai. Each client used the HateLab Platform, now renamed HERO Detect under nisien.ai.

Welsh Government Inclusion and Cohesion Communities Division

The Welsh Government's national strategy 'Prosperity for All' (published September 2017) contains a commitment to work with communities, the voluntary sector and local services to counter the threat of extremism and hate crime in communities. One of our priorities is to work in partnership with the four Welsh police forces and Local Authorities to strengthen our approach to community tension monitoring in Wales. In we aim to:

(a) Balance the short term response (led by police) with long term response (led by Welsh Government, Local Authorities and other partners)

(b) Make best use of the 'word on the street' evidence that can be provided by many partners, and blend this effectively with evidence held by the police from sources including counter-terrorism, police intelligence and social media.

(c) Strengthen multi-agency partnership working to address the immediate and underlying causes of community tensions.

The HateLab Dashboard will provide a valuable contribution to this work. This new source of intelligence will allow our regional community cohesion coordinators to gain insights from social media into emerging community tensions. This will help Welsh Government to deliver on our policy areas of: (i) Community safety and counter extremism; and (ii) Community cohesion and tackling hate crime.

The Welsh Government's Inclusion and Cohesion Team has valued the opportunity to be part of the HateLab Dashboard Pilot. We will outline our approach to using the Dashboard and how we have utilised the system to best suit our needs.

The Welsh Government's Community Cohesion Programme funds a network of eight Community Cohesion teams across eight regions in Wales. As part of their work programme, the Community Cohesion teams monitor community tensions in their respective regions, working with the appropriate partners to mitigate these issues as they arise. The Cohesion Teams provide monthly monitoring reports to Welsh Government but will also send ad hoc updates if the matter is urgent or could potentially escalate.

Monitoring exercises using HERO Detect:

As part of this pilot, we wanted to explore how the HateLab dashboard could potentially help supplement and enhance the work of the Cohesion teams. Since October 2021, whenever we have been alerted to emerging or potential community tensions that we felt appropriate, we have created queries on the Dashboard to track these developing situations for any flashpoints which could result in hateful communications on social media.

Therefore, most of our activity during the pilot has been informed by intelligence supplied by our Cohesion teams. We feel these 'on the ground' sources lend themselves well to the Dashboard, as many of the incidents tended to be localised, with trackable and unique words useful for searching, such as Welsh place names or locations. This has helped to ensure in some cases that we have not collected as much irrelevant information. In some cases, the attached location was one which could be found in multiple countries, sometimes more than one in the same country, such as Newport. With the support from Cardiff University staff, we looked at how to refine our searches.

Regarding the types of incidents we have used the Dashboard to follow, it is difficult to provide a case study as the incidents chosen were highly sensitive and it most cases the incidents are ongoing, whether unresolved or still under investigation. In general, we have used the Dashboard to monitor very localised situations/incidents, which in every case, was not a hate incident, but had the potential to attract negative attention and hateful comments online. This approach has meant many of our queries did not identify hateful comments as the situations thankfully did not escalate to that point. This frequent outcome reflected our intended use of the Dashboard.

For us, the value of the Dashboard was not just in identifying hate speech, but also to monitor where tensions did not result in hate speech online. An example of this would be the migration of Ukrainians to the UK following the Russian invasion in February 2022. We used HateLab to monitor signs of any negative reaction to the news that Ukrainian families would be relocated to Wales, as there was a chance that far-right groups might capitalise on the situation with anti-migrant rhetoric. Concerns were increased following an incident in Hermon, Pembrokeshire where Ukrainian flags were torn down, as reported by the BBC. While we ran these queries, HateLab did not pick up

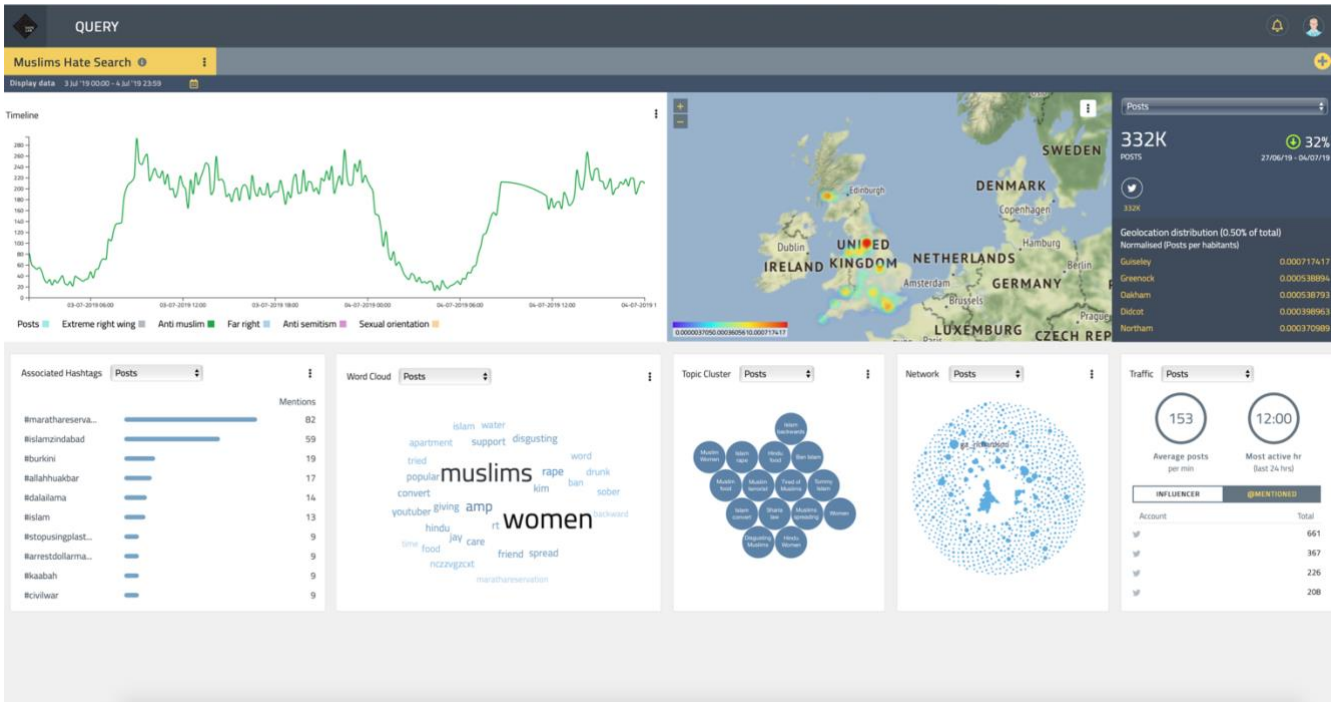
any examples of hate speech online via the queries we set up. This tallied with the general feeling of goodwill towards Ukrainians that was supported by numerous other sources.

Likewise, there was an incident were a far-right group wrongly identified a hotel as a location where asylum seekers would be accommodated. The use of hotels to accommodate asylum seekers is an area which attracts interest from several far-right groups, who use these cases to cause division in the local communities. We used HateLab to monitor whether this misinformation would spread and result in localised tensions or hate speech. Again, it did not, which for us was a positive and provided additional reassurance that the situation had passed without escalation.

In those cases where situations did escalate, the Dashboard showed us how quickly social media accounts from outside the community/area can take an interest in an incident or developing situation. For example, a protest about a particular cause will likely attract other like-minded accounts via hashtags or influencers. In these cases, we saw far more activity as more users were attracted to the developing situation.

As previously mentioned, we are aware that service users have become more savvy in the way they direct abuse, and often do not use openly hateful language, instead choosing to use more coded words. An example would be the use of 'Ok groomer' towards LGBTQ+ people, in particular people that speak in favour of relationship and sexuality education in schools. The search function on HateLab provides a way of homing in on these terms through the lens of the anti-LGBTQ+ classifiers, and provides us with a method of widening our searches and generally allowing us to be more dynamic to developing terms or words.

Going forward, we have considered how the Dashboard could help us to respond with counter messaging/counter narratives. The word cloud function effectively pulls together common phrases, which in most cases are not overly hateful, but can support or promote stereotypical tropes and 'dog whistle' messaging. For example, our queries relating to asylum seekers have picked up the term 'economic migrants' which pushes a trope around people moving to the UK for financial reasons rather than them seeking sanctuary. With some co-ordination with partners, we would like to explore the use of targeted communications to counter such messaging.



The HateLab Dashboard was piloted in Welsh Government's Inclusion and Cohesion Communities Division in 2020-2021.

Galop (LGBTQIA+ anti-violence charity)

For the last 35 years Galop has been working to support the rights of LGBT+ people in the UK, through our work on hate crime, domestic abuse, sexual violence and policing. Our professional casework and helpline services focus on empowerment based support, advice and advocacy to people facing online and offline abuse and violence. We also deliver policy work addressing hate crime by producing research, guidance documents and training. Additionally, we work closely with government, criminal justice, charity and academic bodies to provide advice on issues relating to hate crime. Within the UK we lead the Community Alliance to Combat Hate, an intersectional partnership of leading anti-hate crime charities. Internationally we work closely with the European Commission, OSCE and overseas LGBT+ organisations to stand against online hate speech and hate crime.

Monitoring exercises using HateLab platform:

The HateLab platform was used by Galop for several months, supported by Dr. Arron Cullen. During this period, Galop ran several different monitoring exercises to track anti-LGBT hate online. These exercises were conducted using common and fewer common hashtags being used by those with anti-LGBT+ views. A sampling exercise was conducted using the platform to gather these hashtags, as well as using common hashtags of abuse from Galop's own service users. Once gathered the monitoring exercises were conducted. The exercises allowed Galop to pick up unknown hashtags

that were being used online by those with anti-LGBT+ attitudes, which provided a steppingstone in discovering and tracking a wider set anti-LGBT+ beliefs than was previously possible without the platform.

For example, Galop ran a 'Monkeypox' monitoring exercise for several months using the platform. This exercise allowed us to identify when there were spikes in homophobic and transphobic comments being made online and to match these against offline world events. Using the platform, we saw a spike of homophobic comments against gay men when news articles on the outbreak were trending. This fed into Galop's narrative that there was a strong belief online that 'Monkeypox' was a 'gay' illness, fuelled by certain media outlets, that were helping generate homophobic stigma.

The HateLab platform proved to be a powerful research tool and Galop was very grateful to be part of the pilot. The learning from the pilot showed that the platform can provide strong evidence on online hate against our communities that is often left undetected. Overall, the HateLab platform is a brilliant resource.

National Online Hate Crime Hub

The UK National Online Hate Crime Hub was established by the Home Secretary in 2017 to tackle online forms of hate crime, that increased dramatically in the aftermath of the 2016 referendum vote on the future of the UK in the EU. It acts as the point of contact for all victims of online hate crime, and produces intelligence reports.

Monitoring exercises using HateLab platform:

The provision of the HateLab platform, co-created with the Hub, has fundamentally changed the way we monitor the spread of hate speech during national events. Prior to the platform, Hub staff relied on the Twitter platform interface to gather evidence on the ebb and flow of hate speech around events, such as the referendum vote on the future of the UK in the EU. This proved to be an inadequate method of generating the required insights to track and respond to the problem. During live operations, we were quickly inundated with irrelevant information and failed to capture hate speech in a systematic and reliable way. Through our close collaboration with HateLab, we have co-created technical solutions to overcome these problems. The platform employs sophisticated Machine Learning algorithms to automatically classify hate speech across recognised characteristics at scale and speed, and displays results via a range visualisation tools (frequency chart, top hashtags, topic clusters, geo-location, networks etc). This ensures the Hub can monitor the production and spread of hate speech around events in robust and reliable way.

To date, we have used the platform to monitor hate speech around key moments of the Brexit process, including the abuse of MPs, around terror attacks, and most recently around LGBTQ+ pride month. During these events the platform has allowed the staff in the Hub to better understand the dynamics of hate speech propagation, leading to

improved response times, better support for victims and more effective allocation of resources.

Our latest monitoring exercise around LGBTQ+ pride month in the UK was conducted because our intelligence suggested we would see higher levels of online hate speech during this event than we would typically expect to see throughout the rest of the year. The HateLab platform allowed us to clearly track online hate at the regional and national level during the event. We were able to determine trends on an hourly basis, generate reports centred around keywords and trending hashtags used by those writing the messages, allowing us to create a fuller view of the issue. The platform was an invaluable resource providing a swift analysis of hateful and harmful material in popular discourse allowing us to assess and respond to community threats in real-time.

Overall, the platform provides professionals with the best informed assessment of societal tensions and, combined with nearly a decade of observations of the analytics of social media societal tensions, community intelligence and crime trend data, it enables police managers and partners to make the best-informed decisions on deployment and preventative interventions.

Women's Euros 2022 EE Hope United Campaign

Women's Euros YouTube Video 'Behind the tech - A behind the scenes look at the new EE Hope United shirt': https://www.youtube.com/watch?v=yohzSHrU_x4

HateLab/nisien.ai won contracts to provide hate speech data to EE and British Telecom for their national Hope United Campaigns. For the Women's Euros we provided Saatchi & Saatchi and The Mill with datasets containing hateful and hopeful online messages sent to professional female football players for EE's national anti-misogyny campaign during the tournament. Marketing assets using hero data and insights included apparel, TV, OLV, OOH, and Digital.

"EE has also partnered with HateLab/nisien.ai, a global hub for data and insight into hate speech and crime, to provide each player with their own personalised Hope United shirt. Using behavioural data, which scrapes information from each player's social media account, the shirts show a visual representation of how people are talking about them online, translating emotions associated with hope (such as love, empathy or inspiration) and hate (for example, racial or gender discrimination) into a visual colour and style, creating a unique design for each squad member."

EE Press Release, 4th July 2022

"HateLab/nisien.ai were instrumental in providing the data foundation for our Women's Euros campaign for EE Hope United. We were introduced to them through the data driven shirt design project, working with The Mill Experience; whereby players' individual data histories of online abuse impacted the unique designs of their Hope United kits. The team's tracking of hate across the tournament allowed us to confidently talk about the levels of misogynistic hate our Hope United players received during the Euros and

reflecting that in reactive press, digital out of home and social which ran over the finals weekend. We are working with the team again, on the World Cup, and the data tracked during that period is a creative jumping off point for a current project – orientated around tackling homophobia within the men's game."

Sophie deGraft-Johnson, Business Leader One EE, Saatchi & Saatchi

"Our project required timely and accurate analysis of a high volume Twitter traffic. We needed to understand what all the players of Hope United were experiencing online and to quantify it. HateLab's /nisien.ai's analysis tools made our project possible. They provided us with reliable numbers assessing the anti-black, anti-female and anti-LGBT sentiment in a huge number of tweets and mentions over several months. Using a team with its own machine learning tools and a solid academic background meant our project could stand up against the media scrutiny we expected. On top of all that, Professor Matthew Williams walked us through the complex process and gave us valuable insights behind the numbers and helped us shape our campaign over time. We're grateful for his help and support."

Will MacNeil, Design Director, The Mill

Men's World Cup 2022 EE GayVAR Campaign

Men's World Cup YouTube Video 'EE Hope United introduce GayVAR (featuring Tom Allen & JoeCole)': https://www.youtube.com/watch?v=iVaG0Qi1coA

For the Men's World Cup we provided insights on posts sent to players from the England and Wales teams during the tournament finding that homophobic hate speech was most prevalent. We also advised on the most effective ways to tackle homophobic hatred, which fed into directly into EE's GayVAR campaign.

"And we go again! #EEHopeUnited is back to tackle homophobic hate in football - the most dominant form of online abuse. Odd given the representation of LGBTQ+ talent on the pitch is incredibly low... but of course that's kind of the point. Even the most casual "banter" makes the game feel hostile and less inclusive to players and fans. So we're here in partnership with Football v Homophobia and with Tom Allen, Joe Cole and famous friends to launch GayVAR, calling out homophobia match by match. From our work with Prof. Matt Williams and his team at HateLab/nisien.ai we know how powerful humour can be, avoiding finger-pointing but creating a nudge towards reflection and action. And when homophobia crops up in the game, as ever, EE has the digital skills to help people tackle it.

HateLab/nisien.ai were an essential partner for our Hope United campaigns around the World Cup. Their knowledge of online harms and their award-winning AI was of huge value and gave us depth of understanding and credible insights we couldn't have otherwise achieved. BT and EE would not hesitate work with them on future projects."

Alice Tendler, Group Head of Marketing and Brand, BT Group

Deutsche Telekom #UnHate Campaign

Deutsche Telekom Vimeo Video: '#Unhate the Making of':
https://vimeo.com/759946386

Deutsche Telecom approached HateLab/nisien.ai for assistance with their Mobile World Congress 2022 installation #Unhate, which used an artistic AI to undo hate speech in an immersive environment. HateLab/nisien.ai provided the hate speech for the installation.

"The immersive #UNHATE Experience is an artistic intervention to counteract the omnipresence of hatred in digital spaces. The result is a peaceful statement that raises awareness of the need for a more considerate interaction with each other. With the help of artificial intelligence (AI), real online hate comments are transformed into aesthetic works of art and thus deprived of their toxic effect. The selection of tweets was made in collaboration with the Hatelab/nisien.ai, an initiative of Cardiff University."

Deutsche Telekom Press Release, 28th February, 2022

References:

Alorainy, W. Liu, H., Burnap, P. and Williams, M.L. 2018. The enemy among us': Detecting cyber hate speech with threats-based othering language embeddings. ACM Transactions on the Web 9(4), article number: 39. Available online here: http://orca.cf.ac.uk/120998/

Álvarez-Benjumea, A. and Winter, F. (2018) 'Normative change and culture of hate: An experiment in online environments' *European Sociological Review*, 34:3.

Álvarez-Benjumea, A. and Winter, F. (2020) 'The breakdown of antiracist norms: A natural experiment on hate speech after terrorist attacks' *Proceedings of the National Academy of Sciences*, 117:37.

Bérubé, M., et al. (2020) 'Social media forensics applied to assessment of post–critical incident social reaction: The case of the 2017 Manchester Arena terrorist attack' *Forensic science international*, *313*.

Bevensee, E. and Ross, A.R. (2018) 'The Alt-Right and Global Information *Warfare' IEEE International Conference on Big Data*, IEEE, Seattle, WA, USA.

Bilewicz, M. et al. (2021) 'Artificial intelligence against hate: Intervention reducing verbal aggression in the social network environment' *Aggressive Behavior*, 47:3.

Burke, B. A. et al. (2010) 'Two Decades of Terror Management Theory: A Meta-Analysis of Mortality Salience Research', *Personality and Social Psychological Review* 14.

Burnap, P. and Williams, M. (2015) 'Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modelling for Policy and Decision Making', *Policy & Internet. 7:2*.

Burnap, P. and Williams, M. L. 2016. Us and them: identifying cyber hate on Twitter across multiple protected characteristics. EPJ Data Science. Vol 5(11). Available online here: https://doi.org/10.1140/epjds/s13688-016-0072-6

Burnap, P. and Williams, M.L. 2015. Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modelling for Policy and Decision Making. Policy & Internet. Vol 7:2. Available online here: https://doi.org/10.1002/poi3.85

Carlson, C.R. and Rousselle, H. (2020) 'Report and repeat: Investigating Facebook's hate speech removal process' *First Monday*.

Chan, J. and Bennett Moses, L. (2017), 'Making Sense of Big Data for Security', *British Journal of Criminology*, 57:2.

Chandrasekharan, E. et al. (2017) 'The bag of communities: Identifying abusive behavior online with preexisting internet data' In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 3175–3187.

Cohen, F. et al. (2004) 'Fatal Attraction: The Effects of Mortality Salience on Evaluations of Charismatic, Task-Oriented, and Relationship-Oriented Leaders', *Psychological Science* 15:4.

Cohen, F. et al. (2012) 'Evidence for a Role of Death Thought in American Attitudes toward Symbols of Islam', *Journal of Experimental Social Psychology* 49:3.

Costello et al. (2024) 'Durably reducing conspiracy beliefs through dialogues with AI'. *Science* 385,1814. DOI:10.1126/science.adq1814

Crandal, C. S. and Eshleman, A. (2003), 'A Justification–Suppression Model of the Expression and Experience of Prejudice', *Psychological Bulletin*, 129:4.
Crowley, J. P. (2013) 'Expressive Writing to Cope with Hate Speech: Assessing Psychobiological Stress Recovery and Forgiveness Promotion for Lesbian, Gay, Bisexual, or Queer Victims of Hate Speech' *Human Communication Research*, 40:2.
Cullen, A. and Williams, M. L. (2022) Online Hate Speech Targeting the England Women's Football Team During the UEFA Women's Euro 2022. [Project Report]. HateLab. Available at: https://hatelab.net/wp-content/uploads/2022/08/Online-Hate-Speech-WEURO-2022.pdf
Cullen, A. and Williams, M. L. (2023) Online Hate Speech Targeting the England and Wales Football Teams During the World Cup 2022. [Project Report]. HateLab.
Demos (2015) *Counter-speech: examining content that challenges extremism online*, London, Demos
Demos (2016a) *From Brussels to Brexit: Islamophobia, Xenophobia , Racism and Reports of Hateful Incidents on Twitter* Research, London, Demos
Demos (2016b) *Islamophobia on Twitter: March to July 2016*. London, Demos.
Demos (2017) *Anti-Islamic Hate on Twitter*, London, Demos.
Durán, R. J. (2021) The Economics of Content Moderation: Theory and Experimental Evidence from Hate Speech on Twitter, *SSRN*, Available at SSRN: https://ssrn.com/abstract=4044098
Durán, R. J. et al. (2022) 'The Effect of Content Moderation on Online and Offline Hate: Evidence from Germany's NetzDG', *SSRN*, Available at SSRN 4230296.
Edwards, G. and Rushin, S. (2019), 'The Effect of President Trump's Election on Hate Crimes', *SSRN*, available online at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3102652.
Eichhorn, K. (2001) 'Re-in/citing Linguistic Injuries: Speech Acts, Cyberhate, and the Spatial and Temporal Character of Networked Environments' *Computers and Composition*, 18:2.
Ernst, N. et al. (2017) 'Extreme parties and populism: an analysis of Facebook and Twitter across six countries' *Information, Communication & Society* 20:1.
Fischer-Preßler, D. et al. (2019) 'Collective sense-making in times of crisis: Connecting terror management theory with Twitter user reactions to the Berlin terrorist attack' *Computers in Human Behavior*, 100:1.
Gallacher, J. and Heerdink, M. (2021) 'Mutual radicalisation of opposing extremist groups via the Internet' *PsyArXiv,* Available at: https://psyarxiv.com/dtfc5/
Garland, J. et al. (2022) 'Impact and dynamics of hate and counter speech online' *EPJ Data Science*, 11:1.
Graham, R. (2016) 'Inter-ideological mingling: White extremist ideology entering the mainstream on Twitter' *Sociological Spectrum* 36:4.
Greenawalt, K. (1989) *Speech Crime & the Uses of Language*, Oxford: Oxford University Press.
Greenberg, J. et al. (2014) 'Terror Management Theory and Research: How the Desire for Death Transcendence Drives Our Strivings for Meaning and Significance', in *Advances in Motivation Science*, Vol. 1, ed. Andrew Elliot, New York: Elsevier.
Greig-Midlane, J. (2014) *Changing the Beat? The Impact of Austerity on the Neighbourhood Policing Workforce*, Cardiff University.

Grimmelmann, J. (2015) 'The virtues of moderation', *Yale JL & Tech*, 17:42.
Hambly et al. (2018) *Hate Crime: A Thematic Review of the Current Evidence*, Home Office: London.
Hanes, E. and Machin, S. (2014), 'Hate Crime in the Wake of Terror Attacks: Evidence from 7/7 and 9/11', *Journal of Contemporary Criminal Justice*, 30:2.
Hangartner, D. et al. (2021) 'Empathy-based counterspeech can reduce racist hate speech in a social media field experiment' *Proceedings of the National Academy of Sciences*, 118:50.
Hassan, G., et al. (2022) 'PROTOCOL: Hate online and in traditional media: A systematic review of the evidence for associations or impacts on individuals, audiences, and communities' *Campbell Systematic Reviews*, 18:2.
HateLab (2019) 'A study of cyber hate on Twitter with implications for social media governance strategies', *Conference on Truth and Trust Online*. 4th – 5th October, London, UK.
Hawdon, J. et al. (2017) 'Exposure to Online Hate in Four Nations: A Cross-National Consideration', *Deviant Behavior*, 38:4.
HM Government (2022), *Online Safety Bill*, DCMS, London
HMICFRS (2018) *Understanding the Difference: The Initial Police Response to Hate Crime*. Her Majesty's Inspectorate of Constabulary and Fire and Rescue Service.
Home Office (2022) *Hate Crime, England and Wales, 2021 to 2022*, Home Office: London.
Innes, M. et al. (2018) 'Ten "Rs" of social reaction: Using social media to analyse the "post-event" impacts of the murder of Lee Rigby' *Terrorism and Political Violence*, 30:3.
ISD (2018) Counter Conversation: A model for direct engagement with individuals showing signs of radicalisation, London, Institute for Strategic Dialogue.
Ivandic, R. (2019) 'Jihadi Attacks, Media and Local Hate Crime', *Centre for Economic Performance Discussion Paper 1615*, London: London School of Economics and Political Science.
Jhaver, S. et al. (2019) '"Did You Suspect the Post Would be Removed?" Understanding User Reactions to Content Removals on Reddit', *Proceedings of the ACM on Human-Computer Interaction*, 3:1.
Kaakinen, M et al. (2018b) 'Did the risk of exposure to online hate increase after the November 2015 Paris attacks? A group relations approach' *Computers in Human Behavior*, 78:2.
Kaakinen, M. et al., (2018a) 'Social Capital and Online Hate Production: A Four Country Survey', *Crime, Law and Social Change* 69:4.
Katsaros, M. et al. (2022) 'Reconsidering Tweets: Intervening during Tweet Creation Decreases Offensive Content' *Proceedings of the International AAAI Conference on Web and Social Media*, 16:1.
King, R. D. and Sutton, G. M. (2014), 'High Times for Hate Crimes: Explaining the Temporal Clustering of Hate Motivated Offending', *Criminology*, 51:3.
Kunst, M. et al (2021) 'Do "Good Citizens" fight hate speech online? Effects of solidarity citizenship norms on user responses to hate comments' *Journal of Information Technology & Politics*, *18*:3.

Landau, M. J. et al., 'Deliver Us from Evil: The Effects of Mortality Salience and Reminders of 9/11 on Support for President George W. Bush', *Personality and Social Psychology Bulletin* 30 (2004), 1136–50.
Law Commission (2021) *Hate Crime Laws: Final Report*, London: Law Commission Available at https://www.lawcom.gov.uk/project/hate-crime/
Leets, L. (2001) 'Responses to Internet Hate Sites: Is Speech Too Free in Cyberspace?' *Communication Law and Policy*, 6:2.
Legewie, J. (2013), 'Terrorist Events and Attitudes Toward Immigrants: A Natural Experiment', *American Journal of Sociology*, 118:1.
Liu, H. Burnap, P. Alorainy, W. and Williams, M.L. 2019. Fuzzy multi-task learning for hate speech type identification. Presented at: The Web Conference 2019, San Francisco, CA, USA, 13-17 May 2019WWW '19 The World Wide Web Conference. ACM pp. 3006-3012., Available online here: 10.1145/3308558.3313546
Liu, H., Burnap, P. Alorainy, W. and Williams, M.L. 2019. A fuzzy approach to hate speech classification with two stage training for ambiguous instances. IEEE Transactions on Computational Social Systems 6(2), pp. 227-240. Available online: 10.1109/TCSS.2019.2892037
Liu, Y. et al. (2022) 'Implications of Revenue Models and Technology for Content Moderation Strategies', *Marketing Science*, 41:4.
Menshikova, A. and van Tubergen, F. (2022) 'What Drives Anti-Immigrant Sentiments Online? A Novel Approach Using Twitter' *European Sociological Review*, 38:5.
Müller, K. and Schwarz, C. (2020), 'From Hashtag to Hate Crime: Twitter and Anti-Minority Sentiment' *SSRN*, Available at: https://ssrn.com/abstract=3149103
Müller, K. and Schwarz, C. (2021) 'Fanning the Flames of Hate: Social Media and Hate Crime'. *Journal of the European Economic Association*, 19:2.
Munger, K. (2017) 'Tweetment effects on the tweeted: Experimentally reducing racist harassment' *Political Behavior*, 39:3.
Munger, K. (2021) 'Don't@ me: Experimentally reducing partisan incivility on Twitter' *Journal of Experimental Political Science*, 8:2.
Ofcom (2019) *Adults' Media Use and Attitudes*, Ofcom, London.
Ofcom (2021) *Children and Parents: Media Use and Attitudes*, Ofcom, London.
Ofcom (2022a) *The Online Experiences Tracker (2021/22): Summary Report,* Ofcom, London.
Ofcom (2022b) *Online Nation 2022 Report*, Ofcom, London.
Ozalp, S. et al. (2020) 'Antisemitism on Twitter: Collective Efficacy and the Role of Community Organisations in Challenging Online Hate Speech' *Social Media + Society*, 6:2.
Ozanne, M. et al. (2022) 'Shall AI moderators be made visible? Perception of accountability and trust in moderation systems on social media platforms' *Big Data & Society*, 9:2.
Parekh, B. (2012). 'Is there a case for banning hate speech?' In M. Herz & P. Molnar (Eds.), *The content and context of hate speech: Rethinking regulation and responses*, Cambridge: Cambridge University Press.
Peddell, D. et al. (2016) 'Influences and Vulnerabilities in Radicalised Lone Actor Terrorists: UK Practitioner Perspectives' *International Journal of Police Science and Management* 18:3.

Perry, B. and Olsson, P. (2009) 'Cyberhate: The Globalisation of Hate' *Information & Communications Technology Law*, 18:1.
Roberts, C. et al (2018: 'After Woolwich: Analyzing open source communications to understand the interactive and multi-polar dynamics of the arc of conflict' *The British Journal of Criminology*, 58:2.
Scharwächter, E. and Müller, E. (2020) 'Does terrorism trigger online hate speech? On the association of events and time series' *The Annals of Applied Statistics*, 14:3.
Schmuck, D. et al. (2020) 'Drifting further apart? How exposure to media portrayals of Muslims affects attitude polarization' *Political Psychology*, 41:6.
Scrivens, R. et al. (2021) 'Triggered by defeat or victory? Assessing the impact of presidential election results on extreme right-wing mobilization online' *Deviant Behavior*, 42:5.
Siegel, A. A. and Badaan, V. (2020) '#No2Sectarianism: Experimental approaches to reducing sectarian hate speech online' *American Political Science Review*, 114:3.
Sprejer, L. et al. (2022) 'An actor-based approach to understanding radical right viral tweets in the UK' *Journal of Policing, Intelligence and Counter Terrorism*. DOI: 10.1080/18335330.2022.2086440
Stephan, W. G. and Stephan, C. W. (2000), 'An Integrated Threat Theory of Prejudice', in S. Oskamp, Mahwah, NJ: Erlbaum, eds., *Reducing Prejudice and Discrimination*. Hove: Psychology Press.
Stier, S. et al. (2017) 'When populists become popular: comparing Facebook use by the right-wing movement Pegida and German political *parties' Information, Communication & Society* 20:3.
Sunstein, C. R. (2017) *#Republic: Divided Democracy in the Age of Social Media* Princeton: Princeton University Press.
Tyagi, A. et al. (2020) 'Affective Polarization in Online Climate Change Discourse on Twitter' In Proceedings of *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM),* DOI: 10.1109/ASONAM49781.2020.9381419
Tyler, T. (2019) 'Social media governance: can social media companies motivate voluntary rule following behavior among their users?' *Journal of Experimental Criminology*, 17.
University of Zurich (2025), 'Can AI Change Your View? Evidence from a Large-Scale Online Field Experiment', Available at: https://regmedia.co.uk/2025/04/29/supplied_can_ai_change_your_view.pdf
Vidgen et al. (2019) *How Much Online Abuse Is There: A Systematic Review of Evidence for the UK*, London, The Alan Turing Institute.
Vidgen, B. et al. (2022) 'Islamophobes are not all the same! A study of far right actors on Twitter' *Journal of Policing, Intelligence and Counter Terrorism*, *17*:1.
Williams, M. (2021) *The Science of Hate: How Prejudice Becomes Hate and What We Can Do To Stop It*, London, Faber and Faber.
Williams, M. and Burnap, P. 2016. Cyberhate on social media in the aftermath of Woolwich: A case study in computational criminology and big data. British Journal of Criminology, vol. 56, no. 2, pp. 211–38. Avialble online here: https://doi.org/10.1093/bjc/azv059

Williams, M. and Burnap, P. 2018. Antisemitic Content on Twitter, London: Community Security Trust. Available online here: https://cst.org.uk/public/data/file/4/2/Antisemitic%20Content%20on%20Twitter.pdf

Williams, M. L. a. B., P., Javed, A., Liu, H., and Ozalp, S. (2020) Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime. British Journal of Criminology

Williams, M. L. and Burnap, P. (2016) 'Cyberhate on social media in the aftermath of Woolwich: A case study in computational criminology and big data', *British Journal of Criminology*, 56:2,

Williams, M. L. and Mishcon de Reya (2019) *The Online Hate Speech Report*, London: Mishcon de Reya.

Williams, M. L. and Tregidga, J. (2014) 'Hate Crime Victimization in Wales: Psychological and Physical Impacts Across Seven Hate Crime Victim Types', *The British Journal of Criminology*, 54:5.

Williams, M. L. et al. (2017a) 'Crime sensing with big data: The affordances and limitations of using open-source communications to estimate crime patterns' *The British Journal of Criminology*. 57:2.

Williams, M. L. et al. (2017b) 'Towards an ethical framework for publishing Twitter data in social research: taking into account users' views, online context and algorithmic estimation', *Sociology*, 51:6.

Williams, M. L. et al. (2020) 'Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime', *The British Journal of Criminology*, 60:1.

Williams, M. L. et al. (2022) 'The Effect of the Brexit Vote on the Variation in Race and Religious Hate Crimes in England, Wales, Scotland and Northern Ireland', *The British Journal of Criminology*, https://doi.org/10.1093/bjc/azac071

Williams, M.L., Burnap, P., Sloan, L. (2017) 'Crime sensing with big data: The affordances and limitations of using open-source communications to estimate crime patterns' The British Journal of Criminology. 57(2), 320–340. Available online at: https://doi.org/10.1093/bjc/azw031

Wulczyn, E. (2017) 'Ex machina: Personal attacks seen at scale' In *Proceedings of the 26th international conference on world wide web*, 1391–1399.

Yang, J. and Counts, S. (2010) 'Predicting the speed, scale, and range of information diffusion in Twitter.' In *Proceedings of International conference on weblogs and social media* (ICWSM).

Yin, D. (2009) 'Detection of harassment on web 2.0' In *Proceedings of the Content Analysis in the WEB*, 2: 1–7.