Written evidence submitted by Professor Matthew L. Williams and Dr. Pete Burnap of Cardiff University. Professor Williams is a criminologist in the School of Social Sciences and Dr. Burnap is a Data Scientist in the School of Computer Science and Informatics. Together they Direct the Social Data Science Lab. They are experts on hate and extremism online and they advise the Welsh Government, Home Office, Metropolitan Police Service and Community Security Trust. They have received numerous government grants (Research Councils UK) to study the manifestation, propagation and countering of hate and extremism online. Professor Williams led the UK’s largest independent study of hate crime (All Wales Hate Crime Study) between 2010-13 (funded by the Big Lottery).

Summary:

- Cyberhate (hate speech online) is a scientifically observable and measurable phenomenon. Burnap and Williams (2015) in ‘Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making’, Policy & Internet, evidence the effectiveness of their cyberhate machine learning classifier for race and religion and demonstrate its usefulness for policy and decision making.

- Burnap and Williams (2016) in ‘Us and them: identifying cyber hate on Twitter across multiple protected characteristics’, EPJ Data Science, expand on their classifier to include online hate speech detection for sexual orientation and disability.

- Williams and Burnap (2016) in ‘Cyberhate on social media in the aftermath of Woolwich: A case study in computational criminology and big data’, British Journal of Criminology, show how feelings of heightened threat and a need for revenge are a factor in spikes of hate crime following terrorist attacks that target victims’ way of life – there were observable spikes after 9/11, 7/7, the killing of Lee Rigby and more recent terror attacks.

- Models developed by Williams and Burnap (2015) following the Woolwich attack showed spikes in media reporting correlated with spikes in online hate speech. During the early stages following the Woolwich attack some internet users may be fueled by coverage in the press who have a role in ‘setting the agenda’ and ‘transmitting the images’, especially those users who wish to spread hate, biased rumours and speculation.

- Counter speech online can have a positive effect by stemming the propagation of hate and, when involving groups of people, reinforces norms of acceptable behaviour. Williams and Burnap (2015) show that social media users engaged in counter speech, including the police and media, can stem the spread of hate online.

- Williams, Burnap and Sloan (2016) in ‘Crime sensing with big data: the affordances and limitations of using open source communications to estimate crime patterns’, British Journal of Criminology, show how signals from Twitter data can assist in offline
crime pattern estimation. This methodology is being extended to examine the possible link between online and offline hate.

- Williams M and Tregidga (2014) in ‘Hate Crime Victimisation in Wales: psychological and Physical Impacts Across Seven Hate Crime Victim Types’, British Journal of Criminology, show there is a measurable difference in the experienced impact of hate crimes and incidents for different strands and victims of hate crime, with victims of transgender and disability hate crime reporting suffering the most.

- Groups that are frequently victimised can become immune to the impacts of hate crime, meaning they underreport and are underrepresented in crime surveys.

- The All Wales Hate Crime Project found across all protected characteristics there are a number of factors that victims take into account when deciding to report a hate crime/incident to the police or a third party organization: Whether the offender is known to the victim; whether the incident is an isolated event or part of an ongoing experience; the severity of the incident; and the presence of tangible proof that the incident took place.

- The All Wales Hate Crime Project found alcohol was a factor in hate crime. It was particularly prevalent in hate crimes against members of lesbian, gay, bisexual and transgender (LGBT) communities given nightlife spots were key targets for hate crime offenders.

- The Eyst ‘Think Project’ is a good example of a successful prevention programme that reduced racism and hate offline. It was a 3-year project designed to challenge racism and far right extremism in vulnerable young people in Wales. The project aimed to provide disengaged young people in Wales with greater knowledge and understanding about race, religion, immigration asylum and extremism so that they could challenge and deconstruct racist views.

1. Hate and Extremism Online

1.1 Despite cyberhate being evident from the birth of the domestic Internet (initially with the launch of the Stormfront website in 1995\(^1\)), it has only recently become identified as a social problem that requires addressing. Levin (2002) studied how US right-wing groups promoted their goals on the Web largely unchallenged by law enforcement, concluding that the online medium has been useful to hatemongers because it is economic, far reaching and protected by the First Amendment. Perry and Olsson (2009) found that the Web created a new common space that fostered a ‘collective identity’ for previously fractured hate groups, strengthening their domestic presence in counties such as the US, Germany and Sweden. They warn a ‘global racist subculture’ could emerge if cyberhate is left unchallenged. Eichhorn (2001) focuses on how the online environment opens up the possibility for a more immediate and radical recontextualization of hate speech, while also highlighting its affordances for more effective modes of response, such as vigilantism and counter-speech. Leets (2001) in a study of the impacts of hate related web-pages found that respondents perceived the content of these sites as having an indirect but insidious threat, while Oksanen et al. (2014) show how 67 per cent of 15 to 18 year olds in their study had been exposed to hate material on Facebook and YouTube, with 21 per

---

\(^1\) Stormfront existed in bulletin board format in the early 1990s before being reformed as a website.
cent becoming victims of such material. This final study evidences how the rise of social media platforms has been accompanied by an exponential increase in cyberhate (see also Williams & Wall 2013).

1.2 Open and widely accessible social media technologies, such as Twitter, are increasingly being used by citizens on a global scale to publish content in reaction to world events. The rapid uptake of these technologies has resulted in a massive distributed ‘social sensor net’ that affords government with the opportunity to identify, monitor and trace social reactions to events to the second in real-time. The diffusion of information in these networks following events can manifest itself in a number of ways, ranging from support of social resilience through calls for assistance and advice, to the socially disruptive, through the production and contagion of misinformation and antagonistic and prejudiced commentary (Williams and Burnap 2015).

1.3 Research has shown that the prevalence and severity of crimes with a prejudicial component are influenced in the short term by singular or clusters of events. Acts of terrorism have been shown to influence the prevalence of anti-immigrant sentiment and hate crimes and incidents. On a European scale Legewie (2013) established a significant association between anti-immigrant sentiment and the Bali and Madrid terrorist bombings using Eurobarometer data. Similarly, King and Sutton (2014) found an association between terrorist acts and a rise in hate crime incidents in the US. Convincingly, they show that following the 9/11 terrorist attack law enforcement agencies recorded 481 hate crimes with a specific anti-Islamic motive, with 58 percent of these occurring within two weeks of the attack (4 percent of the at risk period of 12 months). In the UK Hanes and Machin (2014) found significant increases in hate crimes reported to the police in London following 9/11 and 7/7 (28 and 32 per cent increase in the month following respectively). These authors conclude hate crimes cluster in time and tend to increase, sometimes dramatically, in the aftermath of antecedent ‘trigger’ or galvanizing events, such as terrorist acts. They postulate that hate crimes are communicative acts, often provoked by events that incite retribution in the targeted group, towards the group that share similar characteristics to the perpetrators.

1.4 To study the spread of cyberhate on social media the Social Data Science Lab at Cardiff University undertook research funded by the Economic and Social Research Council and Google UK. The project examined the nature of online hate speech across five victim types online: disability; race/ethnicity; religion/belief, sexual orientation, and gender. Data were collected from social media around events that resulted in the production of hateful and antagonistic content (e.g. Woolwich terror attack). Using innovative computational machine learning techniques and statistical modelling Williams and Burnap (2015) were able to automatically classify hateful content and in over 10 million tweets and predict its production and spread following these events. Their findings revealed that hate speech spiked in the first 24 hours following the Woolwich attack, and rapidly deescalated, indicating cyberhate has a half-life (much like offline hate crimes following similar events). Cyberhate was also unlikely to spread to a great extent following the event. Partly explaining this, they found that while information flows from far right political tweeters lasted the longest in the first 24 hours following the event, indicating that this group made use of the attack to further their viewpoint, their
hateful content did not spread beyond a core group of likeminded tweeters. Williams and Burnap (2015) also found that social media information flows from police were the second longest lasting within the first 36 hours after the event, indicating that law enforcement online communications are an effective channel to inform the public and to solicit information.

1.5 Williams and Burnap (2015) found that counter-speech was a common response to extremism or hateful content online. Extreme posts are often met with disagreement, insults, and counter-speech campaigns. Combating hate speech with counter speech has some advantages over government and police responses: it more rapid, more adaptable to the situation and pervasive; it can be used by any internet user (e.g. members of the public, charities, the media, the police); and it draws on nodal governance and responsibilisation trends currently prominent in the wider criminal justice system. The following typology of counter hate speech was identified:

- Attribution of Prejudice
  e.g. “Shame on #EDL racists for taking advantage of this situation”
- Claims making and appeals to reason
  e.g. “This has nothing to do with Islam, not all Muslims are terrorists!”
- Request for information and evidence
  e.g. “How does this have anything to do with the colour of someone's skin??”
- Insults
  e.g. “There are some cowardly racists out there!”

Initial evidence from ongoing experiments with social media data show that counter speech is effective in stemming the length of hateful threads when multiple unique counter speech contributors engage with the hate speech producer. However, not all counter speech is productive, and evidence shows that individuals that use insults against hate speech producers often inflame the situation, resulting in the production of further hate speech.

1.6 Williams, Burnap and Sloan (2016) show how crime and disorder-related posts on Twitter are associated with actual police recorded crime rates offline. They evidence that: i) social media communications can be mined for mentions of crime and disorder; ii) these mentions can be statistically linked to offline police crime data; iii) online communications about crime and disorder are positively associated with offline patterns in particular areas in London; and iv) where biases exist in online communications these can be identified and adjusted for to generate reliable estimates of offline crime patterns. This study has shown that naturally occurring social media communications have the potential to provide an alternative information source on the offline hate crime problem.

2. All Wales Hate Crime Project:

2.1 As the Crime Survey for England and Wales cannot be reliably extrapolated to smaller constituent regions in isolation when studying hate crime (Office for National Statistics 2012) the All Wales Hate Crime Project provided the first look at the hate crime problem in Wales, allowing the Welsh Government to better target resources and evaluate
the effectiveness of their interventions. The Project was the largest and most comprehensive study of hate crime in the UK, surveying circa 2000 members of the public from minority groups and interviewing over 60 victims.

2.2 The Project provided the first multi-victim-type analysis of hate crime, showing that psychological and physical impacts of crimes and incidents are not homogenous across the seven victim types. Neither are all hate crime types equal in their subsequent negative impacts upon victims. The study identified physical and psychological impacts that were predicted by victim type, crime type and demographic factors:

- Nearly a fifth (18%) of respondents attempted to conceal their identity post-victimisation
- Nearly a third (29%) of victims had thoughts about moving from their local area
- Nearly one in five (18%) had considered moving out of Wales entirely
- One in seven hate crime victims reported having suicidal thoughts
- Victims of repeat victimisation were over four times more likely than any other victim to experience thoughts of suicide
- Being unemployed and having a negative ‘sense of belonging’ to a local area increased the likelihood of suffering multiple impacts
- Repeat victimisation (by the same offender) was by far the strongest predictor of multiple impact experience
- Transphobic hate crime victims were most likely to suffer a high number of psychological and physical impacts, followed by victims of disability hate crime
- Victims of violent hate crimes were most likely to suffer a high number of psychological and physical impacts, compared to victims of other types of hate crime

2.3 The Project found across all protected characteristics there are a number of factors that victims take into account when deciding to report a hate crime/incident to the police or a third party organization: Whether the offender is known to the victim; whether the incident is an isolated event or part of an ongoing experience; the severity of the incident; and the presence of tangible proof that the incident took place. A large number of interview participants highlight inconsistencies in reporting and recording mechanisms, and there are examples where hate-identified incidents have been recorded as neighbour nuisance or anti-social behaviour. Almost half (44%) of the survey victim respondents stated they had reported their most serious hate crime to the police. The reasons given for reporting include: victim belief that ‘it was the right thing to do’ (69%); victim desire to ‘stop it happening again’ (62%), and victim hope that the offender would be brought to justice (52%). The reasons given for not reporting include victim belief that: the incident was ‘too trivial’ (29%); the police could not have done anything (27%); the incident was a private matter that could be dealt with personally (19%). The vast majority of victims had not been put off contacting the police and said they would encourage other hate crime victims to do so. However, there were a number of issues raised with regards to the third-party (independent non-police) reporting systems in operation in Wales. Currently, the majority of third party reporting options are perceived to be online and it is evident that this often serves as a barrier for some people who do not have access to IT facilities.
3 Recommendations for action by the Government

3.1 Most online hate speech research has been conducted on social media platforms (mainly Twitter and Facebook). There is a need to examine hate on emerging platforms and online gaming.

3.2 Further research is needed to look into the relationship between hate speech online and offline hate crime.

3.3 There is very limited evidence on how the police deal with reports of online hate crime and incidents. A Home Office study should be commissioned to conduct an audit at force level. This audit should evaluate the existing reporting mechanisms and how police record the modus operandi of hate (on or offline or a mix).

3.4 Questions relating to online hate victimisation should be included in the Crime Survey for England and Wales.
References


