Institution: Cardiff University		
Unit of Assessment: Sociology (21)		
Title of case study: HateLab: Preventing the rising tide of hate crime and speech through influencing policy and policing		
Period when the underpinning research was undertaken: 2011 – 2020		
Details of staff conducting the underpinning research from the submitting unit:		
Name(s): Williams, Matthew	**Role(s) (e.g. job title):** Professor	**Period(s) employed by submitting HEI:** 01/10/2001 – present
Period when the claimed impact occurred: 01/10/2013 – 31/12/2020		
Is this case study continued from a case study submitted in 2014? No		

1. Summary of the impact (indicative maximum 100 words)

Police-recorded hate crimes in England and Wales are at their highest levels since records began, and the rise in online hate requires the police to address the problem both offline and online. Cardiff's ESRC-funded HateLab is the first to address the problem on both fronts, generating vital evidence to inform policy and operational decisions on prevalence, impact and prevention. This evidence formed the primary source of information for the Welsh Government's Framework for Action on Tackling Hate Crime (2014). Cardiff HateLab technologies were also embedded within the National Cyber Hate Crime Hub (run by the UK's National Police Chiefs' Council), allowing policymakers and police to address hate crime and speech.

2. Underpinning research (indicative maximum 500 words)

A significant 'dark figure' remains in hate crime victimisation estimates. Police recorded hate crime data, which indicate a ~150% increase over the last decade, are inaccurate due in part to volatile changes in reporting and recording practices. The Crime Survey for England and Wales (CSEW) provides more reliable hate crime estimates, showing a downward trend over the last decade, but it also suffers from methodological limitations rendering annual and regional (e.g. Wales) estimates moot. The disparity between sources and unreliable regional estimates has hindered policy development. The migration of hate crime to the internet also created additional operational and policy needs not currently met by existing sources of information.

Williams is Director at Cardiff University's HateLab, a global hub for data and insight into hate crime and speech. Research within HateLab draws on a range of methods, from conventional surveys to machine learning, and an array of criminological concepts, from moral panics to hierarchies of crime information credibility, to generate a broad evidence base on hate offline and online within Wales, the UK and beyond.

The All Wales Hate Crime Project (AWHCP) **[G3.1]**, the first project of its kind in the UK, was the foundation for the HateLab research programme. It remains the largest, most comprehensive study of hate crime in the UK, surveying ~20,00 members of the public, interviewing ~60 victims and engaging with ~5,000 citizens **[3.1]**. This mixed method approach examined the impacts of hate crime across protected characteristics (race, religion, sexual orientation, disability, and transgender identity), as well as age and gender.

Key findings include **[3.1]**:

- First insight into the psychological and physical impacts of hate crime across the seven victim-types (most previous hate crime research had focused on discrete victim-types in isolation). Those suffering transgender and disability motivated hate crimes were by far the most likely to suffer multiple types of impact compared to other victim types;

- Hate crime can have considerable physical and/or psychological impacts on victims (and their families), for example post-victimisation, 18% of respondents attempted to conceal their identity, 18% had considered moving out of Wales, and one in seven experienced suicidal thoughts;
- Evidence on victims' experiences of the criminal justice process that improved service delivery;
- Inconsistencies in reporting and recording mechanisms, as well as issues with independent non-police reporting systems in Wales – many of these are perceived to be online only which acts as a barrier to victims without access to IT facilities;
- A large proportion of victims highlighted considerable challenges to accessing support and, as a result, 'suffered in silence';
- Evidence of the emerging trend in online hate speech on Facebook and Twitter.

The project provided a robust source of survey data on hate crime not previously available to policy-makers or practitioners in Wales **[3.1]**.

2.1 Identifying and monitoring online hate speech

Four projects **[**~£3M from ESRC and US Department of Justice, **G3.2]** were funded between 2013-2020 with extensive policy cooperation. The then Head of the Cross-Government Hate Crime Programme, and now NPCC National Policing Advisor for Hate Crime, acted as a co-investigator on two projects. These projects addressed the challenges of identifying, monitoring and stemming online hate speech. Williams combined statistical modeling techniques with collaborator Burnap's (Cardiff University) computer science machine learning approach to cope with the unprecedented scale and speed of social media data that had hitherto been a barrier to social science researchers. This novel interdisciplinary approach generated the first evidence base on online hate in the UK **[3.2, 3.3, 3.4]**.

For example, the first criminological analysis of online social reactions to terror attacks in the UK (2014, 2015) revealed that anti-Muslim hate speech spiked in the first 24 hours, and rapidly deescalated, indicating a 'half-life' of 'cyberhate'. Police online information flows were the second longest lasting within the first 36 hours after these events, indicating these as effective channels to keep the public informed, solicit information, and counter hate speech **[3.2, 3.3, 3.4]**. Subsequent analysis showed online hate speech positively correlates with offline hate crimes on the streets of London (2019) **[3.5]**. Further key findings and outcomes from these funded projects include:

- Survey and social media data showing hate crime and online hate speech spiked in the final weeks of the leave campaigns and following the Brexit vote **[3.5]**;
- Statistical models that showed online anti-Muslim and anti-black hate speech can predict race and religious hate crime offline **[3.5]**;
- Evidence that counter-hate narratives delivered by third sector actors are effective in stemming the production of hate speech **[3.6]**;
- Development of the HateLab Dashboard – the first platform to use Machine Learning (ML) to detect online hate speech in real time and at scale – to assist academics, practitioners and policy makers in monitoring the spread of online hate speech **[3.3]**. The development of ML to detect hate speech drew heavily on social science concepts, such as Membership Categorisation Analysis, the process of 'othering' in language, perceived threat and dehumanisation;
- Analysis of the ethical issues associated with using ML to detect and monitor hate speech in real-time, which contributed to the ongoing debate on predictive policing **[3.2, 3.3, 3.5]**.

3. References to the research (indicative maximum of six references)

[3.1] Williams, M. L. and Tregidga, J. (2014) Hate crime victimisation in Wales: psychological and physical impacts across seven hate crime victim-types. *British Journal of Criminology* 54(5), pp. 946-967, DOI: 10.1093/bjc/azu043

[3.2] Williams, M. L. and Burnap, P. (2015) Cyberhate on social media in the aftermath of Woolwich: A case study in computational criminology and big data. *British Journal of Criminology* 56(2), pp. 211-238, DOI: 10.1093/bjc/azv059

[3.3] Burnap, P. and Williams, M. L. (2015) Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet* 7(2), pp. 223-242, DOI: 10.1002/poi3.85

[3.4] Burnap, P. and Williams, M. L. et al. (2014) Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack. *Social Network Analysis and Mining* 4, pp. 206, DOI: 10.1007/s13278-014-0206-4

[3.5] Williams, M. L., Burnap, P., Lui, H., Javed, A. and Ozalp, S. (2019) 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,* 60 (1) , pp. 93-117, DOI: 10.1093/bjc/azz049

[3.6] Ozalp, S., **Williams, M. L., Burnap**, **P**., Liu, H. and Mostafa, M. (2020) Antisemitism on Twitter: collective efficacy and the role of community organisations in challenging online hate speech, *Social Media and Society* 6(2), pp. 1-20, DOI: 10.1177/2056305120916850

Selected grants:

[G3.1] All Wales Hate Crime Project, Big Lottery Fund, £569,194 (2011-2013)

[G3.2] HateLab, ESRC/US DoJ, ~£3M total funding over 4 projects:

- HateLab: Real-Time Scalable Methods & Infrastructure for Modelling the Spread of Cyberhate on Social Media, ESRC, £1,842,478 (2017-2022)
- Hate Crime After Brexit: Linking Terrestrial and New Forms of Data to Inform Governance, ESRC, £249,995 (2019-2021)
- Understanding Online Hate Speech as a Motivator for Hate Crime, National Institute for Justice, US DoJ, $885,820 (2016-2019)
- Hate Speech and Social Media: Understanding Users, Networks and Information Flows, ESRC/Google, £124,986 (2013-2014)

4. Details of the impact (indicative maximum 750 words)

Williams' work provided a more accurate picture of a) the impacts of hate crime in Wales and b) hate speech online. The novel evidence provided, unavailable from other sources, underpinned the Welsh Government's Framework for Action on Hate Crime and provided an effective way for UK police forces to respond to online hate. This allowed policymakers and practitioners to better identify, monitor, and prevent hate.

4.1 Creating a new framework for hate crime in Wales

Prior to Williams' work, the Welsh Government did not have robust data on hate crime in Wales, as existing data sources – such as the Crime Survey for England and Wales – could not be reliably extrapolated to devolved administrations. Cardiff's All Wales Hate Crime Project (AWHCP) was foundational to the Welsh Government's hate crime policy through directly informing the national Framework for Action on Tackling Hate Crime **[5.1]** – the first of its kind in Wales. The Framework focuses on prevention, supporting victims and improving the multi-agency approach. Underpinned by Williams' research, it cites the AWHCP more than any other research source (25 times). John Davies, the Welsh Government's Head of Inclusion and Cohesion, confirmed that the Cardiff team *"directly [fed] in robust data and insights from their research"* and that the AWHCP *"provided a much needed novel source of evidence"*

[5.2]. He cited three examples of evidence provided which were *“essential to developing key parts of the Framework”* **[5.2]**:

- Evidence on the impacts of hate crime on victims living in Wales across protected characteristics;
- Victims’ experiences of the criminal justice system;
- The emerging trend in online hate speech.

An evaluation carried out by the Welsh Government Audit Office (2015-16) stated *“the Framework was…supported by robust evidence from the All Wales Hate Crime Project…. [it] helped…the Framework to focus upon areas which will have a significant impact”* **[5.3**, p.1**]**. For example, under-reporting and the importance of community reporting systems were key findings of the AWHCP that fed into the Framework Delivery Plan. Welsh Government funding was allocated as a result to improve third-party (independent non-police) systems for victims to report hate crimes. Davies confirmed that *“a major innovation of the Framework, supported by Professor Williams’ evidence that shows significant barriers exist to reporting hate crime in Wales, is the Wales Hate Crime Report and Support Centre”* **[5.2]**. This Centre provides a flexible reporting mechanism, enhanced and integrated support for victims, and has a dedicated team of case workers. Davies highlighted an evaluation which found that the Centre contributed to an increase in hate crime reporting of 20%, despite victimisation falling by 28% in the same period **[5.2]**.

4.2 Supporting delivery of policy actions and influencing practice

Williams is the only academic member of the All Wales Hate Crime Criminal Justice Board that takes forward the Framework’s recommendations. His ongoing work draws on HateLab research **[3.3, 3.5, 3.6]** to deliver the following actions from the Framework Delivery Plan (2015-6) **[5.4]**:

- Pilot a Community Tension Detection System to assess community feeling on social media during high periods of tension (the HateLab Dashboard);
- Conduct research on: i) the possible link between online hate speech and offline hate crimes and ii) identifying and modelling the spread of anti-Semitic content on Twitter.

In addition to informing policy, All Wales Hate Crime Project findings *“have influenced practice and public engagement via Welsh Government outputs for end-users”* **[5.2]**. Outputs developed using the findings include:

- Online Hate Speech Guides for use in schools, as well as by youth groups and probation services; these have been accessed 6,120 times **[5.5a** and **5.5b]**;
- Campaign materials for National Hate Crime Awareness Weeks (2013-2017) featuring films of victims of hate crime surveyed by the AWHCP speaking of their experiences, overlaid with highlights of the research findings **[5.6]**;
- A Tackling Hate Crime Toolkit, which was distributed to all Housing Associations in Wales, has been used by a range of practitioners (including housing officers responsible for dealing with hate crime, managers, and contractors) to support training on estates with high levels of hate crime. The toolkit directly quotes issues highlighted in the AWHCP and also quotes a key project recommendation: *“Social housing providers need to ensure that hate crime offenders are dealt with quickly and effectively and policies do not result in processes that ‘manage’ the victim rather than deal with the perpetrators (All Wales Hate Crime Research Project)”* **[5.7**, p.12**]**.

At UK level, the project has been widely quoted in national policy documents, including ‘Action Against Hate: The UK Government’s Plan for Tackling Hate Crime’ (2016) **[5.8a]**. Burnap was also invited, as a HateLab representative, to the **Home Affairs Select Committee’s inquiry on Hate Crime and its Violent Consequences** in 2016, set-up in response to the murder of Jo Cox MP and rising levels of hate speech and crimes. HateLab evidence was cited in the committee’s report showing that online hate speech could be detected at scale and in real-time with the ML developed at Cardiff **[3.2 - 3.5]**. The inquiry's subsequent criticisms of social media companies for not using such ML to counter the spread of hate influenced the drafting of the White Paper on Online Harms **[5.8b]**.

4.3 Tackling online hate speech for the UK's National Cyber Hate Crime Hub

Police statistics on online hate significantly underestimate the extent of the problem as most incidents go unreported. The HateLab Dashboard, by monitoring in real time and not having to wait for victims to file a report, reflects a direct observation of prevalence. It was designed to facilitate near-real time responses to cyberhate, including targeting counter-speech at repeat perpetrators and pre-empting outbreaks of hate crime on the streets. It is currently used by Welsh Government Community Cohesion Coordinators, the Wales Extremism and Counter Terrorism Unit, and a London-based LGBTQ anti-violence charity **[5.9]**.

The Dashboard has also been integrated into the UK-wide National Cyber Hate Crime Hub (2019), run by the National Police Chiefs' Council. The Hub, the point of contact for all victims of online hate crime, uses the Dashboard to produce intelligence reports for police, senior civil servants and MPs. Paul Giannasi, Director of the Hub, is responsible for coordinating policy and operational activity across all Government departments and criminal justice agencies in their response to hate crime. He confirmed that the new data and insights provided by the research *"transformed the way my team of police officers and civilian staff at the National Cyber Hate Crime Hub monitor and tackle online hate crime"* and that the Dashboard has *"fundamentally changed the way we monitor the spread of hate speech during national events"* **[5.10]**.

Prior to the Dashboard, Hub staff did not have an adequate way of tracking and responding to hate speech around events, such as the Brexit referendum vote. Giannasi stated: *"during live operations, we were quickly inundated with irrelevant information and failed to capture hate speech in a systematic and reliable way"* but the technological solutions provided by HateLab *"ensures the Hub can monitor the production and spread of hate speech around events in a robust and reliable way"* **[5.10]**.

Dashboard data on the spread of online hate speech allows the police to better understand the dynamics of propagation, leading to improved response times, better support for victims and more effective allocation of resources **[5.10]**. Most recently, *"the Dashboard has allowed us to gain valuable insight into how groups and individuals have used the coronavirus outbreak to stir up hatred online"* **[5.10]**.

Giannasi estimated that HateLab has resulted in economic savings of ~£500,000 via the Dashboard, cloud and data services and an implementation evaluation **[5.10]**.

5. Sources to corroborate the impact (indicative maximum of 10 references)

[5.1] Tackling Hate Crimes and Incidents: A Framework for Action 2014

[5.2] Testimonial: John Davies, the Welsh Government's Head of Inclusion and Cohesion

[5.3] Tackling Hate Crimes and Incidents: Framework for Action Progress Report 2015-16

[5.4] Hate Crime Criminal Justice Board Cymru - Delivery Plan 2015-16, 2016-17, 2017-18

[5.5] **a.** Online Hate Speech Guides **b.** Website analytics

[5.6] National Hate Crime Awareness Week materials

[5.7] Hate Crime and Housing Toolkit

[5.8] **a.** Home Office: Action Against Hate: The UK Government's plan for tackling hate crime; **b.** Hate Crime and its Violent Consequences, Home Affairs Select Committee, Oral Evidence II (Dr Pete Burnap)

[5.9] ESRC grant proposal, including evidence of Dashboard use

[5.10] Testimonial: Paul Giannasi, Director of the National Cyber Hate Crime Hub