QUEEN'S UNIVERSITY BELFAST

Offensive communications: exploring the challenges involved in policing social media

Williams, M., Butler, M., Jurek-Loughrey, A., & Sezer, S. (2021). Offensive communications: exploring the challenges involved in policing social media. *Contemporary Social Science*, *16*(2), 227-240. https://doi.org/10.1080/21582041.2018.1563305

Published in:
Contemporary Social Science

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Publisher rights
© 2019 Academy of Social Sciences. This work is made available online in accordance with the publisher's policies. Please refer to any applicable terms of use of the publisher.

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.

Download date:06. Apr. 2021

Contemporary Social Science

Themed Issue on 'Evidence Policy in a Digital Society'

Guest editors: Linda Hantrais (corresponding guest editor), Aygen Kurt-Dixon and Ashley Thomas Lenihan

Offensive communications: Exploring the challenges involved in policing social media

Authors:

Mark Williams (Corresponding author), Centre for Secure Information Technology (CSIT), Queen's University Belfast, Northern Ireland Science Park, Queen's Road, Queen's Island, Belfast, BT3 9DT, United Kingdom. Email: mwilliams03@qub.ac.uk, Tel: 07557388471

Michelle Butler, School of Social Sciences, Education & Social Work, Queen's University Belfast, Belfast, United Kingdom.

Anna Jurek-Loughrey, School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, United Kingdom.

Sakir Sezer, The Centre for Secure Information Technology (CSIT), Queen's University Belfast, Belfast, United Kingdom.

Notes on contributors

Mark Williams is a Leverhulme Interdisciplinary Network on Cybersecurity and Society (LINCS) postgraduate research student at the Centre for Secure Information Technologies (CSIT) at Queen's University Belfast (2016–2019). His research explores the interface between the social sciences and electronic engineering and computer science focussing primarily on the criminal use of social media. In his project, he is looking at ways of detecting and preventing inappropriate and criminal behaviour in social media, with particular emphasis on the definition, classification and mitigation of offensive online expressions.

ORCID: 0000-0002-8159-1900

Michelle Butler is a Lecturer in Criminology at Queen's University Belfast. She graduated from the University of Kent with a Master's Degree in Forensic Psychology and a PhD in Criminology from the University of Cambridge. Her research interests include criminological psychology, penology, and the management of crime.

ORCID: 0000-0002-6983-6215 Twitter: @MichelleBQUB

Anna Jurek-Loughrey is a Lecturer in the School of Electronics, Electrical Engineering and Computer Science at Queen's University Belfast. She graduated from the Technical University of Lodz in Poland with Masters Degrees in Computer Science and Applied Mathematics. She obtained her PhD in Computer Science from Ulster University. Her work has spanned a diverse set of topics comprising: machine learning; sensor-based activity recognition within smart environments; social media analytics with application in security and public health; sentiment analysis; record linkage and entity resolution.

Sakir Sezer is Professor of Secure Digital Systems at Queen's University Belfast, where he leads the Networked System Security at the Centre for Secure Information Technology. He is an international expert in high-performance network processing and internet security technologies, spanning cybersecurity-related topics in malware, embedded systems, IoT, ICS and network security. His main research focus has been on the acceleration of network security-related functions by exploring novel hardware-based parallel processing architectures, System on Chip (SoC), NPU/NFP (Netronome) and programmable technologies, including MPSoC and FPGA.

Acknowledgements

The research reported in this article was carried out with funding provided by a postgraduate research scholarship from the Leverhulme Interdisciplinary Network on Cybersecurity and Society (Project code: R2179PAI) from September 2016–2019.

Word length 6,985

Offensive communications: Exploring the challenges involved in policing social media

Abstract

The digital revolution has transformed the potential reach and impact of criminal behaviour. Not only has it changed how people commit crimes but it has also created opportunities for new types of crimes to occur. Policymakers and criminal justice institutions have struggled to keep pace with technological innovation and its impact on criminality. Criminal law and justice, as well as investigative and prosecution procedures, are often outdated and ill-suited to this type of criminality as a result. While technological solutions are being developed to detect and prevent digitally-enabled crimes, generic solutions are often unable to address the needs of criminal justice professionals and policymakers. Focussing specifically on social media, this article offers an exploratory investigation of the strengths and weaknesses of the current approach used to police offensive communications online. Drawing on twenty in-depth interviews with key criminal justice professionals in the United Kingdom, the authors discuss the substantial international challenges facing those seeking to police offensive social media content. They argue for greater cooperation between policymakers, social science and technology researchers to develop workable, innovative solutions to these challenges, and greater use of evidence to inform policy and practice.

Keywords: digital policing; social media; offensiveness; offensive content online; technological policing interventions.

The rise of social media has revolutionised public communication (Burnap & Williams, 2016; Hanson, 2016). However, its growth has also contributed to the committing of new crimes and provided new avenues within which to commit more traditional crimes (House of Lords, 2014). In particular, its ability to reach a wide audience in a short timeframe has raised concerns about its use to spread hate crime, extremism and disinformation, as well as abusive, threatening and offensive content (Citron, 2014; Awan, 2016; Home Affairs Committee, 2017). In this regard, the policing of social media should be seen in the wider context of the regulation of the internet more generally and the role that criminal justice organisations, users and platform providers are supposed to play in preventing crime online (Trottier, 2012; Stenning & Shearing, 2015; Murray, 2016; Gillespie, 2018; Yar, 2018). Increasingly, governments are calling on social media companies and social media users to do more to police these networks and reduce the harm they cause (European Union Internet Forum, 2016). However, policing social media has often proven difficult due to the following factors: the volume of the number of posts that need to be policed; the inter-jurisdictional nature of users; the lack of international cooperation and information-sharing protocols; the ease and anonymity by which the content can be disseminated; and varying legal definitions regarding what qualifies as a crime and warrants prosecution (Council of Europe, 2003; Wall, 2013). Nowhere is this more obvious than when attempting to police offensive content on social media (Mangan & Gillies, 2017).

Research has consistently highlighted problems defining what constitutes offensiveness, since what is considered to be offensive varies across time and place, depending on the social, cultural, political and religious norms in different jurisdictions and even within different social groups within the same jurisdiction (Frank, 2001; Jay & Janschewitz, 2008; Gray & Ford, 2013; *The Economist*, 2018). These variations may affect judgements regarding what social media postings it is in the public interest to prosecute, contributing to disparities in policing policies and practices (Laidlaw, 2017). For example, in the United States users are unlikely to be prosecuted for posting abusive and threatening social media content due to cultural, social and political norms emphasising the freedom of speech (Elonis v. United States, 135 S. Ct. 2001 (2015)). Yet, in the United Kingdom (UK) users can and are regularly prosecuted for posting similar content under Section 127 of the Communications Act 2003 (Ministry of Justice, 2016). Moreover, challenges are involved in assessing the extent to which users intended to cause offense, especially if their posts have been shared without their consent to a larger audience than intended (Scaife, 2013).

Despite international agreement that online abuse and hate are a growing problem, no international consensus or political urgency to harmonise legislation is evident (Rohlfing, 2015), resulting in difficulties in the prevention, investigation and detection of this type of behaviour (Giannasi, 2015). Often, current international approaches to policing offensiveness on social media tend to rely on the police and/or social media companies manually reviewing content to judge its offensiveness, making these processes onerous, inefficient and expensive (Gillespie, 2018). Policing of social media content therefore tends to be retrospective, with victims reporting such behaviour to the police and/or social media companies after harm has already occurred (Awan & Zempi, 2016).

Police officers are increasingly under pressure to do more to regulate social media in a ‘space that is privately owned but publicly populated’ (Wall, 2013, p. 455). This demand for action is due to victims’ fears that the offensive and threatening content online will materialise offline and result in their physical harm (Awan & Zempi, 2015, 2017). However, as Trottier (2012) notes, policing social media cannot rely solely on police ‘watching over’ users but must also involve self and peer policing, as well as social media companies regulating content. Social media companies have tended to present themselves as a means to enable communication rather than publishers of content (Kiss & Arthur, 2013). As a result, they argue that their responsibility is one of protecting users from offense or harm that they have experienced, rather than proactively protecting users from exposure to such harm (Gillespie, 2018). This attitude has contributed to a retrospective approach to policing, resulting in the adoption of a 'report and takedown' model in which they do not proactively moderate illegal content but rely on users reporting offensive content for review and possible removal (Laidlaw, 2017; Home Affairs Committee, 2017).

This article draws on a series of interviews with key criminal justice professionals in the UK to examine the strengths and weaknesses of the current approach used to police offensive content online. The authors seek to identify potential enablers and inhibiters of an effective evidence-based approach to the regulation of social media posts, and suggestions are offered for how policymakers, social scientists and computer scientists can work together to develop innovative solutions to ensure the consistent and effective application of government policy in this area.

Researching the policing of offensive communications

A qualitative approach using interviews was employed to obtain a better understanding of what types of offensive online communications are prosecuted, what evidence is needed to acquire convictions, and the challenges involved in working in this area. Interviews were utilised to help uncover the experiences of participants and to elucidate how policies and practices are implemented in institutions (Mason, 2002). In-depth, semi-structured interviews lasting approximately 45 minutes were conducted with 20 professionals in Northern Ireland

and England, including senior police officers, state prosecutors, the judiciary, legal experts and policymakers.

A purposeful sampling approach was used to recruit participants, who were identified based on their professional criminal justice and/or policymaking expertise in this area. Initial contact was made via email, and snowball sampling was used to identify further participants. The voluntary nature of the research was outlined, and participants were notified that they were free to withdraw from the study at any time. Participants were informed that only limited confidentiality and/or anonymity could be offered given the small number of people working in this field. They were also informed that any disclosure of serious criminality, harm to oneself/another or staff malpractice would be reported to a relevant authority. Ethical approval for the research was obtained from authors' institution.

The interview data were coded using NVivo, and thematic analysis was used to identify, analyse and record recurring patterns or themes in the interview data, concerning the participants' understanding of offensive content, and to uncover how policy is interpreted and enacted in the detection and prevention of this behaviour (Braun & Clarke, 2006). Anonymised quotes were chosen and presented in the findings to exemplify the themes under discussion. Inter-rater reliability, in which data are independently coded and compared for agreement, was used to ensure that the themes emerging from the interviews were accurate and supported by the underlying data (Armstrong, Gosling, Weinman, & Marteau, 1997). Using Cohen's Kappa coefficient, inter-rater agreement was calculated at 84%. Cohen Kappa's above 75% are usually considered to be very good.

Research findings

The findings from our study are divided into three sections. The first section examines the responsibility of social media companies for dealing with offensive content, followed by a section reviewing the strengths of existing policy and practice in this area and describing potential enablers in the proactive policing of social media content. The third section explores the weaknesses of current policy and practice and identifies potential inhibiters to an evidence-based approach to proactive policing of social media.

The responsibility of social media companies for offensive content

The retrospective ‘report and take down’ model adopted by social media companies is in many respects a response to the European Electronic Commerce Directive 2000/31/EC, which states that social media companies are ‘mere conduits’ for the transmission of information and not liable for its content, on the condition that they do not select or modify the information contained within it (European Parliament and Council, 2000; Mangan & Gillies, 2017). The Court of Justice of the European Union (CJEU) validates this approach and affords protection for social media companies from liability in relation to illegal content, provided they have no knowledge of its content and act expeditiously to remove it when informed (Google France SARL and Google Inc. v Louis Vuitton Malletier SA, 2010, C-236/08 to C-238/08). The voluntary Code of Conduct agreed between the major social media companies, the EU and member states’ hate speech legislation reflects the same approach. The Code commits social media platforms to assessing and removing content that potentially breaches EU rules within 24 hours of user notification (European Union Internet Forum, 2016).

While this approach has seen 70% of reported content in the EU being removed within 24 hours, social media companies continue to apply rules relating to offensive, abusive and hateful content without transparency and inconsistently, usually only acting in response to commercial or media pressure (Home Affairs Committee, 2017; Statt, 2017; European Commission, 2018). Nonetheless, social media companies are being forced to review this approach in light of US and UK investigations into Facebook's privacy practices following the Cambridge Analytica data scandal, in which approximately 87 million Facebook users' data were improperly used allegedly to influence the US Presidential Election and Brexit vote in 2016 (Cadwalladr & Graham-Harrison, 2018). Currently, the lack of openness on the part of social media companies regarding the resources and policies employed to tackle offensive content has resulted in the development of a hierarchy of service. Content reported by high-profile users, the media or commercial partners receives closer scrutiny and more rapid resolution in direct contrast to the perfunctory response experienced by most ordinary users. This hierarchy of service has led to growing frustration on the part of governments and calls for social media companies to play a more proactive role in policing offensive behaviour (Home Affairs Committee, 2017).

Governments are also realising that, due to the scale and global nature of social media, offline-policing practices are difficult to replicate online, and national governments alone cannot protect citizens from online harms (Cohen, 2017; HM Government, 2017; Gollatz, Beer, & Katzenbach, 2018). Discussions regarding responses to the harms users face have, consequently, broadened into a more expansive examination of the responsibilities of social media companies (Home Affairs Committee, 2017; Gillespie, 2018; Oltermann, 2018). This debate has resulted in recent German legislation imposing fines on social media companies who fail to remove illegal content when notified, and similar calls by the UK Home Affairs Committee to introduce fines (Home Affairs Committee, 2017; BBC, 2018). In 2018, the European Commissioner for the Digital Single Market called for social media companies to remove offensive and illegal content proactively and automatically (Lomas, 2018). Gillespie (2018, pp. 4–5) argues that such changes require a paradigmatic shift and a move away from viewing the moderation of offensive content as ‘occasional and ancillary’ to that which is an ‘essential, constant and definitional’ component of what social media companies do.

While the Electronic Commerce Directive does not impose a general obligation on social media companies to moderate content and, in many respects, disincentivises moderation due to the risk of being treated as a publisher, specific exceptions in cases of child pornography and terrorism have been made (Twentieth Century Fox and others v British Telecommunications plc , 2011; The Electronic Commerce Directive (Terrorism Act 2006) Regulations 2007). The distinction between general and specific moderation needs to be explored further in response to offensive content that is clearly unlawful and in cases where a ‘diligent economic operator should have identified the illegality’ (L'Oréal and others v eBay, 2011, p. 7). Recital 48 of the Electronic Commerce Directive provides EU member states with the possibility of imposing on social media companies an expectation to engage with them in the detection and prevention of illegal activities (European Parliament and Council, 2000). HM Government (2017) maintains that, while the illegality of offensive content is potentially problematic, a working model could be achieved through closer cooperation between social media companies and national legislators in defining what should be moderated and how technology can be best used as an enabling device. This approach is echoed by Scaife (2013) who notes that social media companies require further governmental guidance to ensure that rights are respected and a balance is found between user privacy and protection from offensive content.

Reviewing existing policy and practice

One of the key enablers in the successful policing and prosecution of crimes committed via social media in the UK has been the Government's commitment to tackling hate crime, extremism, disinformation, abusive, threatening and offensive content on social media, especially in the wake of the murder of Member of Parliament Jo Cox by a right-wing terrorist over her political views on Brexit during the 2017 election campaign. The commitment to tackling behaviour on social media is evidenced in several UK government publications (for example Home Affairs Committee, 2017; Committee on Standards in Public Life, 2017; HM Government, 2017; Law Commission, 2018).

Reviews are useful for ensuring problems are fully understood and adequate responses coordinated. The move towards greater consistency of understanding, approach and response by the Government is also evidenced in the announcement that a national police online hate-crime hub would be established as a central nexus through which all reports of online hate crime would be channelled (HM Government, 2017). The centralisation and streamlining of such a service aims to ensure that specially trained police officers are available to provide specialist support to local officers investigating online hate, the collection of evidence and provision of support to victims. This focus on consistency and coordination is especially important in the UK given that the devolution of policing to Scotland and Northern Ireland, as well as the 43 territorial forces engaged in policing in England and Wales, provides ample opportunities for deviations in policy and practice to occur.

The Metropolitan Police Service has already used a centrally coordinated, specialist police online hate hub as part of the Mayor's Office for Policing and Crime hate crime reduction strategy. This strategy relies on drawing together 'Met resources alongside partners from social media companies, academic and data analysis experts, and national voices' (Participant

9) to help identify, prevent and investigate hate crime and abuse occurring on Twitter and Facebook. This approach allows officers to police online hate more effectively by developing expertise in this area, and improving the ability of professionals to understand, identify and deal with this issue. As the same participant noted, the Hub facilitates:

> …[an] understanding of where and when and how hate crime is perpetrated online … and move[s] it into a place where we are starting to build the tools that will help us to address it and help the police to understand how they can address it. (Participant 9)

Another participant explained that centralisation helps to ensure a consistent approach, reducing the potential for discretion to result in crimes not being followed up or actioned due to a lack of knowledge or unfamiliarity about what to do when such crimes occur:

> [social media] crimes have a lower chance of being screened out, because by and large the crimes are [normally] looked at by a DS [Detective Sergeant], who goes 'No apparent leads, screen it out. It is a waste of our time. We are not going to go with it'. We [in the Hub] can do better than that. (Participant 11)

The prosecution service works with the hub to build expertise in examining offensive social media content to determine if a prosecution should be brought. As one participant explained:

> The hub [was] set up to establish best practice on how to deal with these cases. How to identify if it meets the threshold for charging and how to deal with attribution. So, it is establishing best practice that could hopefully be disseminated around the country. (Participant 19)

In this way, centralisation was believed to help to ensure consistency in the decision-making process regarding when to prosecute and build expertise in balancing freedom of speech and censorship concerns with the need to prosecute offensive, illegal content:

> It [offensiveness] is also very difficult to define and the fear is that it can very easily be used as a tool for censorship and a massive breach of freedom of speech. …so because

> of that the attempt has been made to have a certain amount of consistency by having the same team always give authority [to prosecute] (Participant 20).

It was hoped that a move towards a consistent approach to investigating and prosecuting these crimes would enable the development of expert knowledge required to investigate, gather and prosecute social media crimes, as well as ensure that examples of both good and bad practice are shared and learned from, rather than lost, forgotten or not used to enhance professionals' skillsets.

Of particular importance is the potential for this centralised, specialist approach to help build knowledge and expertise in undertaking evidential collection for social media crimes. For example, the online hate crime hub in London supports local police officers by 'the capturing of evidence … [and] quality assur[ing] what then happens out at a borough level' (Participant 9). This support is essential for local police officers as it provides 'a package of information that identifies a suspect…[and] help[s] them to identify where that suspect lives' (Participant 11). Ensuring that evidence is collected appropriately and managed properly also aids the successful prosecution of crimes by reducing the potential for evidence to be deemed inadmissible or tainted.

Additionally, the development of centralised expertise helps to ensure that victims can be supported in a more consistent manner, with referrals being made to specialist victim support agencies, as one specialist officer commented:

> You are a victim. We see that you have been insulted by this…Do you mind if we refer you to victim support in the form of Stop Hate UK? They have a particularly good relationship with the social media providers and may well be able to assist you. (Participant 11)

Several participants reported that victims: ‘want the abuse, the harassment, the thing that is causing them distress to stop. In our world that is generally by having entries removed from social media’ (Participant 12). In contrast to local officers, who have not been trained to deal with these crimes and are unaware of the avenues by which victims can request the removal of social media content, specialist officers can easily provide support: ‘victims want the material taken down…that [is] fairly easy for us to achieve’ (Participant 11).

Consequently, the UK government’s commitment to reviewing the laws surrounding offensive online communications and developing a centralised, coordinated approach to policing and prosecution of social media crimes could be expected to aid consistency, the use of evidence-informed policy and practice, as well as the collation of best practice examples that can be disseminated across the different police organisations throughout the UK. Currently, the lack of such an approach is one of the biggest weaknesses of existing policy and practice as knowledge regarding best practice is not being sufficiently collated, analysed or disseminated, resulting in it residing with a few specialist police officers.

Weaknesses of existing policy and practice

A key weakness in the current approach to policing social media is the disparate and disorganised manner in which these crimes are currently dealt with due to a lack of clear guidance and training. The devolution of policing to Scotland and Northern Ireland, as well as the 43 different territorial forces in operation in England and Wales, significantly complicates the provision of a common approach, clear guidance and training when different policies and practices are being adopted and pursued in these various organisations. Apart from the few police officers who had been specially trained in how to respond to these crimes, ordinary police officers reported confusion over what to do in such cases and how evidence should be gathered. They were fearful of making mistakes which could undermine the case against the alleged offender. According to one officer:

> We need to put better guidance out to officers on how they actually respond to these incidents, what the expectations are from the service, what the expectations are from the Police Ombudsman and everyone else, because people are very confused as to actually what they can do and what they can't do; be that capturing the evidence, be that interpreting the evidence, be that asking for help from social media providers and companies…there is a lot of confusion out there. (Participant 8)

Given the prevalence of social media in society, the lack of training on this issue was seen as inhibiting best practice and leaving officers unsure as to what they should do and how they should support victims. Another officer told us:

> We had absolutely no training. You are trained in offences that can occur over social media, but you are not trained [in evidential capture of SM messages] … you are basically told to seize the phone and give it to DESU [Digital Evidence Support Unit]. (Participant 7)

A failure to capture and learn from police officers' experiences in this area weakened the adoption of an evidence-based best practice approach to policing social media content as officers' successes, failures and lessons learned were not being captured, shared or discussed with colleagues or between police organisations in an organised manner. This resulted in an inconsistent approach being taken in the policing of social media crimes since, in the absence of training and clear guidance, officers began to use their discretion in deciding how cases should be responded to and what cases should be sent forward for prosecution, as reiterated by one officer:

> …we don't always have to send prosecution files [for prosecution] … if they have no offending history for that offence, and they admit the guilt… then it can basically be closed at that point. … That was the way that that was dealt with. So, there was no real investigation. (Participant 4)

The use of this discretion is troubling, however. If cases are responded to and dealt with differently by different officers, in different locations or different organisations, the result may be inconsistency of approach, potential discrimination in how different cases and offenders are dealt with, and the undermining of public confidence in policing. These concerns were shared by one of the participants who feared that police discretion could result in the biased delivery of justice as not all cases were being referred to the prosecution service, potentially resulting in a filtering of cases by the police based on the nature of the offence and/or characteristics of the offender/victim:

> We wouldn't encourage that approach [police discretion]. …If there is an offence disclosed, the police have a duty to investigate it. …I acknowledge that there can be difficulties in investigating and prosecuting these cases, but where the evidential test and prosecution and public interest tests are met, really those things shouldn't get in the way. (Participant 5)

The lack of training and a cohesive approach to the investigation of offensive social media communications was also evident in the inadequate resources available to local police officers to capture and process social media content and shortcomings in the collection of digital evidence that local police officers experienced. Outside of the hubs, the police lacked front-line capability to triage digital platforms and capture evidence in a timely, efficient and secure manner. The process was cumbersome, slow and potentially offers perpetrators the opportunity to destroy evidence (by deleting social media content) before devices are collected for evidential purposes using forensic examination in accordance with the UK Association of Chief Police Officers Good Practice Guide for Computer-Based Electronic Evidence (ACPO, 2014). However, this is a lengthy process as there is ‘a seven, eight-month backlog on phones [being triaged]’ (Participant 7). To prosecute such cases successfully, digital evidence must be captured by appropriately trained officers on a secure police terminal. The lack of trained officers and their tendency to work only during regular office hours hindered the speedy capture of evidence, investigation and prosecution of cases: Participant 6 maintained that officers trained to capture digital evidence only worked 9am to 5pm whereas ‘by the very nature of it, we work 24/7’.

These delays allow ample time for evidence to be lost/destroyed and for victims to continue to receive abusive messages. Officers are required to liaise with other specialist officers to connect the suspect’s IP address to a physical address or physical mobile device and then bank records, such as direct debit details, which are usually applied for to prove that a suspect is using the particular device to send offensive social media content to the victim. This process ‘takes about eight weeks or so, which is quite time consuming’ (Participant 6). To overcome delays, avoid perpetrators attempting to destroy evidence and reducing the likelihood of no further action being taken due to lack of evidence, some police officers screenshot offensive messages, although this method of capturing data is not legally admissible: ‘Well I knew that it is not technically allowed, but if it is the only way that we have, I just ran with it, knowing that I was probably going to get told off about it later on’ (Participant 7). Police officers justified screenshotting evidence by the delays in the current process, which they argued hindered efficient prosecution as well as allowing perpetrators to destroy evidence and/or continue to harass victims:

> The thing then with doing everything so officially is that you are not… you don’t have the opportunity of getting him remanded overnight because it is going to take so long to get everything done. So your hands are tied and if that’s the only thing you have to get the evidence… (Participant 7)

The potential, however, exists for evidence collected in this way to be legally challenged or tampered with, as one participant pointed out with regard to a case s/he investigated:

> I was very surprised that they [the suspect] did plead [guilty] to it [harassment, threats to kill and improper use of a public electronic communications network], because if they made an issue about it [how the evidence was captured] they could have got off with it. (Participant 7)

The lack of clear policies, training and current weaknesses in the investigation of social media crimes is worrying not only due to its potential impact on the investigation and prosecution of cases that are reported to the police but also because it may deter potential victims from coming forward to report crimes:

> We know…. that when some people report offences to the police, the police might not recognise them as offences or know what to do in terms of attribution and so on…. So … if the police were trained up to a different level we could have far, far more cases coming into our system to deal with. (Participant 19)

The impact of under-reporting and the lack of consistency in responding to cases when they are reported inhibit the investigation and prosecution of social media crimes in a number of ways. It impedes police understanding of the nature of the offences being committed, the scale of the problem and what resources are needed to tackle such crimes adequately. We found a great sense of frustration amongst investigators and prosecutors that current practices do not lend themselves easily to the effective policing of social media content:

> People … hide behind personas and social media providers, and shout abuse and groom, threat, harass, stalk and commit all those offences without any expectation that the police will come and detect them. And it is very difficult for police to do that. (Participant 8)

The need to invest in technology to aid police investigations was reiterated by a number of participants:

> We need then to invest in technology … with technology we can issue fixed penalties electronically, we can get emails electronically, but we can't do evidential captures. And we need to … have front line capability to grab that when it is required. And we need kiosk examinations to triage mobile phones so that we are not taking very expensive devices off victims for long periods of time. And it needs to be: plug it in a kiosk, download it, evidential content has been grabbed, and there is your phone back. And those kiosk solutions would be in every station. And that would be the norm in the future. Yes, I've got an abusive message on social media. No problem. Plug it in. we have downloaded it, captured the content and given it back. (Participant 8)

Nonetheless, even with these investments it was noted that police may still be overwhelmed due to the prevalence of offensiveness online. Participants expressed the need for a more proactive approach aimed at prevention to be adopted rather than the retrospective approach currently being used:

> I think the fundamental problem is the sheer wealth of material. That's what I think is the main obstacle here … which is why I think one of our views is prevention rather than prosecution. (Participant 20)

These findings point to the dominance of an unstructured, decentralised approach to policing offensive social media content in the UK, resulting in offensive crimes committed online being dealt with differently depending on the discretion and knowledge of local police officers. Existing practice and policy were largely found to be hampering the adoption of a consistent, evidence-based approach to policing social media crimes due to the failure to collate, analyse and disseminate examples of successes and failures to local police officers who were at the frontline in responding to these cases.

Conclusion

While these findings are drawn from evidence gathered in Northern Ireland and England, potentially limiting their generalisability to other jurisdictions, a number of general conclusions can be drawn. Firstly, police forces prefer prevention to prosecution due to the sheer scale of offensiveness online. Even with the development of a centralised, coordinated response, the criminal justice system faces being overwhelmed by the volume of cases. Secondly, closer cooperation between social media companies, legislators and police officers is required to agree what types of offensive online communications will be prosecuted and how a balance between user privacy and protection from offensiveness will be achieved. Thirdly, the development of a centralised, coordinated expertise in policing social media content is a necessary first step to ensure consistency and the development of evidence-based best practices. In particular, specific attention needs to be paid to ensuring the different police and crime commissioners and chief constables –with responsibility for overseeing and developing policy and practice for the 43 territorial forces in England and Wales, the PSNI and Police Scotland – are coordinating and liaising with each other to avoid deviations in policy and practice emerging. Specifically, good practice and lessons from the 'hub' could be disseminated to police officers around the country through the ‘operations’ and/or ‘equality, diversity and human rights' coordinating committees of the National Police Chiefs' Council (NPCC). Fourthly, local police officers need to be better trained and resourced to avoid the potential for discretion to be used in an inconsistent and potentially discriminatory manner, undermining public confidence and police legitimacy. Training can also help improve the investigative and evidence gathering techniques employed by officers, and lessen the potential for evidence to be deemed inadmissible or tainted. Moreover, local police knowledge, which is key for providing insights into contextual factors specific to that area/population, needs to be captured. For instance, police officers in Northern Ireland may be more aware of offensive sectarian messages than their counterparts elsewhere in the UK.

Fifthly, while such a centralised approach would be beneficial, it will still be retrospective in nature. Technology can be used to help develop a more preventative, automatic and proactive model of policing social media content. Such an approach is preferable as it may reduce the publication of offensive material and hence victimisation. Technology offers the possibility to identify potentially criminal content proactively, by combining technological, criminological and legal expertise to ensure that what is censured is not arbitrary but derived through careful analysis, informed by legislation. While the necessary technology is not yet available, the authors are currently working on developing such software using these findings to inform its development and design. For example, machine learning can be used to identify potentially offensive messages as they are being written and a pop-up to inform users of the potential for harm and/or prosecution if the material is posted. Such an approach provides an opportunity for self-censorship, as well as helping to demonstrate intent to cause harm in a prosecution if the pop-up is dismissed. This software can also be used to aid the police in evidence capture by encoding metadata, such as the location of posting, time of posting, demographic details, required for investigation and prosecution. In this way, the investigation of cases could be speeded up, evidence capture could be streamlined and the need to retain personal items, such as mobile phones and computers, for long periods of time could be reduced. By ensuring that experts in computer science, social science and the criminal justice system work together to agree a common understanding of what constitutes offensiveness that is criminally liable, this software can be modified for different jurisdictions based on relevant local legislation, rather than attempting to implement a one-size-fits-all model as employed in most existing technical solutions currently in use in which offensiveness is arbitrarily defined. However, it should be noted that the development of such a technological innovation will be insufficient to address all concerns in this area if it is not also combined with wider criminological research, policy and training in the policing of cybercrimes (see Wall & Williams, 2013; Yar, 2018). Lastly, legislative change is required before social media companies can be prosecuted for not removing liable content of which they have been made aware.

Our findings show that, by building upon current good practice, by developing better platforms and combining the expert knowledge of policymakers, social scientists and computer scientists, mechanisms can be devised to police offensive content on social media proactively. Although our interviews were carried out in Northern Ireland and England, the evidence collected suggests that the mechanisms identified could serve to aid the investigation and prosecution of such cases in other jurisdictions, thereby enhancing public confidence in the criminal justice system and its legitimacy.

References

Association of Chief Police Officers (2014). *Good practice guide for computer-based electronic evidence* (Version 5). Retrieved from https://www.app.college.police.uk/app-content/investigations/forensics/#digital-forensics

Armstrong, D., Gosling, A., Weinman, J., & Marteau, T. (1997). The place of inter-rater reliability in qualitative research: an empirical study. *Sociology*, *31*(3), 597-606. doi:10.1177/0038038597031003015

Awan, I., & Zempi, I. (2015). *We fear for our lives: Offline and online experiences of anti-Muslim hostility*. Retrieved from https://tellmamauk.org/fear-lives-offline-online-experiences-anti-muslim-hostility/

Awan, I. (2016). Islamophobia on social media: A qualitative analysis of the Facebook's walls of hate. *International Journal of Cyber Criminology*, *10*(1), 1-20. doi:10.5281/zenodo.58517

Awan, I., & Zempi, I. (2016). The affinity between online and offline anti-Muslim hate crime: Dynamics and impacts. *Aggression and Violent Behavior*, 27, 1-8

Awan, I., & Zempi, I. (2017). 'I will blow your face off'—Virtual and physical world anti-Muslim hate crime. *The British Journal of Criminology, 57*(2), 362-380.

BBC (2018, January 1). *Germany starts enforcing hate speech law*. Retrieved from http://www.bbc.co.uk/news/technology-42510868

BBC (2018, April 22). *Jeremy Hunt threatens social media with new child-protection laws.* Retrieved from http://www.bbc.co.uk/news/uk-43853678

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77-101. doi:10.1191/1478088706qp063oa

Burnap, P., & Williams, M. (2016). Us and them: Identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, *5*(1), 1-15. doi:10.1140/epjds/s13688-016-0072-6

Cadwalladr, C., & Graham-Harrison, E. (2018, March 19). Facebook and Cambridge Analytica face mounting pressure over data scandal. *The Guardian*. Retrieved from https://www.theguardian.com/news/2018/mar/18/cambridge-analytica-and-facebook-accused-of-misleading-mps-over-data-breach

Citron, D. K. (2014). *Hate crimes in cyberspace*. Cambridge: Harvard University Press.

Cohen, Y. (2017). *The Net is Closing –birth of the e-police*. London: CreateSpace.

Committee on Standards in Public Life. (2017). *Intimidation in Public Life Review*. Retrieved from https://www.gov.uk/government/publications/intimidation-in-public-life-a-review-by-the-committee-on-standards-in-public-life

Council of Europe. (2003). *Convention on cybercrime protocol on xenophobia and racism*. Retrieved from https://edoc.coe.int/en/cybercrime/6559-convention-on-cybercrime-protocol-on-xenophobia-and-racism.html

Doward, J. (2017, October 14). Government's new online hate crime hub given just £200,000. *The Guardian*. Retrieved from https://www.theguardian.com/society/2017/oct/14/government-criticised-for-low-funding-level-to-tackle-online-hate

The Economist (2018, January 13) Germany is silencing 'hate speech' but cannot define it. Retrieved from https://www.economist.com/news/europe/21734410-new-social-media-law-causing-disquiet-germany-silencing-hate-speech-cannot-define-it

European Commission. (2018). *Code of conduct on countering illegal hate speech online. Results of the 3rd monitoring exercise*. Retrieved from http://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=612086

European Parliament and Council. (2000). *Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (Directive on electronic commerce).* Brussels: European Parliament and Council.

European Union Internet Forum. (2016). *Code of conduct on countering illegal hate speech online*. Brussels: European Union Internet Forum.

Frank, M. J. (2001). The social context variable in hostile environment litigation. *Notre Dame Law Review*, *77*, 437-440.

Giannasi, P. (2015). Policing and hate crime. In N. Hall, A. Corb, P. Giannasi, & J. Grieve (Eds), *The Routledge international handbook on hate crime* (pp. 105-116). London: Routledge.

Gillespie, T. (2018, February 6). Moderation is the commodity. *Techdirt.* Retrieved from techdirt.com https://www.techdirt.com/articles/20180206/09403839164/moderation-is-commodity.shtml

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale: Yale University Press.

Gollatz, K., Beer, F., & Katzenbach, C. (2018). *The turn to artificial intelligence in governing communication online*. Retrieved from https://nbn-resolving.org/urn:nbn:de:0168-ssoar-59528-6

Gray, J. A., & Ford, T. E. (2013). The role of social context in the interpretation of sexist humor. *Humor*, *26*(2), 277-293.

Hanson, J. (2016). *The social media revolution: An economic encyclopedia of friending, following, texting, and connecting*. California: Greenwood.

Hinduja, S. (2007). Computer crime investigations in the United States: Leveraging knowledge from the past to address the future. *International Journal of Cyber Criminology*, *1*(1), 1-26. doi: 10.5281/zenodo.18275

HM Government. (2017). *Internet safety strategy green paper*. Retrieved from https://www.gov.uk/government/consultations/internet-safety-strategy-green-paper

Home Affairs Committee. (2017). *Hate crime: Abuse, hate and extremism online*. Retrieved from https://publications.parliament.uk/pa/cm201617/cmselect/cmhaff/609/60902.htm

Home Office. (2017, October 8). *Home Secretary announces new national online hate crime hub* [Press Release]. Retrieved from https://www.gov.uk/government/news/home-secretary-announces-new-national-online-hate-crime-hub

House of Lords' Select Committee on Communications (2014). *Social media and criminal offences*. Retrieved from https://publications.parliament.uk/pa/ld201415/ldselect/ldcomuni/37/3703.htm

Jay, T., & Janschewitz, J. (2008). The pragmatics of swearing. *Journal of Politeness Research. Language, Behaviour, Culture,* 267-88. doi:10.1515/JPLR.2008.013

Kiss, J., & Arthur, C. (2013, July 29). Publishers or platforms? Media giants may be forced to choose. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2013/jul/29/twitter-urged-responsible-online-abuse

Laidlaw, E. (2017). What is a joke? Mapping the path of a speech complaint on social networks. In D. Mangan & L. Gillies (Eds). *The legal challenges of social media* (pp. 127-154). Cheltenham: Edward Elgar.

Law Commission. (2018) *Offensive online communications*. Retrieved from https://www.lawcom.gov.uk/project/offensive-online-communications/

Lomas, N. (2018, January 9). Europe keeps up the pressure on social media over illegal content takedowns. *Tech Crunch*. Retrieved from https://techcrunch.com/2018/01/09/europe-keeps-up-the-pressure-on-social-media-over-illegal-content-takedowns/

Mangan, D., & Gillies, L. (2017). *The legal challenges of social media*. Cheltenham: Edward Elgar.

Mason, J. (2002). *Qualitative researching*. London: Sage.

Mayor of London (2017, April 24). *Mayor launches new unit to tackle online hate crime* [Press Release]. Retrieved from https://www.london.gov.uk/press-releases/mayoral/mayor-launches-unit-to-tackle-online-hate-crime

Ministry of Justice. (2016). *Criminal Justice System Statistics Quarterly*. December 2016. Retrieved from https://www.gov.uk/government/statistics/criminal-justice-system-statistics-quarterly-december-2016

Murray, A.D. (2017) Mapping the rule of law for the internet. In D. Mangan & L. Gillies (Eds). *The legal challenges of social media* (pp. 13-36). Cheltenham: Edward Elgar.

NCA Strategic Cyber Industry Group. (2016). *Cybercrime assessment*. London: National Crime Agency (NCA).

Oltermann, P. (2018, January 5). Tough new German law puts tech firms and free speech in spotlight. *The Guardian*. Retrieved from https://www.theguardian.com/world/2018/jan/05/tough-new-german-law-puts-tech-firms-and-free-speech-in-spotlight

Rohlfing, S. (2015). Hate on the internet. In N. Hall, A. Corb, P. Giannasi, & J. Grieve (Eds), *The Routledge international handbook on hate crime* (pp. 293-305). London: Routledge.

Scaife, L. (2013). The interrelationship of platform providers and users in the regulation of Twitter and offensive speech: Is there a right to be offensive and offended at content? *Communications Law*, 18(4), 128 - 134.

Statt, N. (2017, March 24). YouTube is facing a full-scale advertising boycott over hate speech. *The Verge*. Retrieved from https://www.theverge.com/2017/3/24/15053990/google-youtube-advertising-boycott-hate-speech

Stenning, P. & Shearing, C. (2015). Privatisation, pluralisation and the globalisation of policing. *Research Focus, 3*(1), 1 – 8.

Suzor, N. (2018). Digital constitutionalism: Using the rule of law to evaluate the legitimacy of governance by platforms. *Social Media & Society*, *4*(3), 1-11.

Trottier, D. (2012) Policing social media. *Canadian Review of Sociology,* 49.4, 411-425

Wall, D.S. & Williams, M.L. (2013) Policing cybercrime: Networked and social media technologies and the challenges for policing, *Policing and Society*, *23*(4), 409-412, doi: 10.1080/10439463.2013.780222

Wall, D.S.(2013) Policing identity crimes, *Policing and Society*, *23*(4), 437-460, doi: 10.1080/10439463.2013.780224

Yar, M. (2018). A failure to regulate? The demands and dilemmas of tackling illegal content and behaviour on social media. *International Journal of Cybersecurity Intelligence & Cybercrime*, 1(1), 5-20.