
How much online abuse is there?

A systematic review of
evidence for the UK

Policy Briefing – Full Report

Bertie Vidgen, Helen Margetts,
Alex Harris

**The
Alan Turing
Institute**

Public Policy Programme
Hate Speech: Measures
and Counter Measures

Contents

Authors	2
Introduction	3
Key findings	4
Recommendations.....	7
Background.....	8
Defining abuse.....	10
The prevalence of online abuse in the UK.....	12
1. UK Government statistics on illegal online abuse	12
2. Civil Society reports of online abuse	16
3. Platforms' transparency reports on abuse.....	18
4. Measurement studies	28
5. Survey evidence on experiences of online abuse	33
Acknowledgements	46
References	46
Research publications and reports	46
News articles, blogs and websites	52

Funding

This work was supported by Wave 1 of The UKRI Strategic Priorities Fund under the EPSRC Grant EP/T001569/1, particularly the “Criminal Justice System” theme within that grant, and The Alan Turing Institute.



**UK Research
and Innovation**

Authors

Bertie Vidgen is a post-doctoral researcher at The Alan Turing Institute, a Research Associate at the University of Oxford and Visiting Fellow at the Open University

www.turing.ac.uk/people/researchers/bertie-vidgen

Helen Margetts is the Director of the Public Policy Programme at The Alan Turing Institute and Professor of Society and the Internet at the Oxford Internet Institute, University of Oxford

<https://www.turing.ac.uk/people/programme-directors/helen-margetts>

Alex Harris is a Research Assistant in the Public Policy Programme at The Alan Turing Institute

www.turing.ac.uk/people/researchers/alexander-harris

Introduction

Online abuse, which includes both interpersonal attacks, such as harassment and bullying, and verbal attacks against groups (usually called 'hate speech'), is receiving more attention in the UK (HM Government 2019; SELMA 2019; The Law Commission 2018). It poses myriad problems, including inflicting harm on victims who are targeted, creating a sense of fear and exclusion amongst their communities, eroding trust in the host platforms, toxifying public discourse and motivating other forms of extremist and hateful behaviour through a cycle of 'cumulative extremism' (Eatwell 2006).

Understanding the prevalence of online abuse is crucial for addressing more complex and nuanced issues, such as what its causes are, when and where it manifests, what its impact on society is and how we can challenge it. The Home Office and Local Communities secretaries captured this point in 2018: 'Hate crime is a complex issue [...]. In order to tackle it, we need to understand the scale and nature of the problem, as well as the evidence about what works in tackling it.' (Home Office 2016). At a time when the UK Government is considering greater regulation of online harms, building an appropriate evidence base is key. However, to date relatively little attention has been paid to this fundamental question: *How much online abuse is there?*

Part of the challenge is that, at present, the data, tools, processes and systems needed to effectively and accurately monitor online abuse are not fully available and the field is beset with terminological, methodological, legal and theoretical challenges (Brown 2018; Davidson et al. 2019; Vidgen et al. 2019). And, despite the hype about computational tools for the automated monitoring of online behaviour, algorithms alone will not resolve the challenge of how to best detect and measure online abuse (Ofcom 2019c). As Facebook CEO Mark Zuckerberg reported during the 2018 American Senate hearings on disinformation, 'Hate speech – I am optimistic that, over a 5 to 10-year period, we will have AI tools that can get into some of the nuances [...] But, today, we're just not there.' (The Washington Post 2018)

In this policy briefing paper from The Alan Turing Institute's [Hate Speech: Measures and counter-measures](#) project, we estimate the prevalence of online abuse within the UK by reviewing evidence from five sources: (i) UK Government figures, (ii) reports from civil society groups, (iii) transparency reports from platforms, (iv) measurement studies, primarily from academics and thinktanks and (v) survey data. We also present previously unpublished results from the Oxford Internet Survey (OxIS) 2019. In some cases, UK-specific evidence cannot be attained and evidence from other countries or global reports are used, which is flagged where needed.

Key findings

1. The available evidence is fragmented, incomplete and inadequate for understanding the prevalence of online abuse. Appropriate statistics are difficult to find and, in many cases, are not provided with the necessary contextual information to fully interpret them. For instance, some of the big platforms share how much abusive content they have removed – but not how much content they host in total.
2. The prevalence of legally defined online abuse is incredibly low. 1,605 online hate crimes were recorded in England and Wales in 2017/18 and 1,067 in 2016/17. Across all types of online abuse, which includes online harassment, we estimate that there is fewer than 1 offence per 1,000 people in the UK. Noticeably, in the Home Office's 2018/19 hate crime report, no figures were given for online hate due to concerns about the quality of statistics. This is a serious limitation of existing reporting.
3. The prevalence of online abuse on mainstream platforms which is serious enough for them to action is also very low. We estimate that it is ~0.001%, although this figure is inherently speculative because platforms do not share how much total content they host. However, we note that concerns have been raised about whether platforms moderate sufficiently, with some critics suggesting they leave substantial amounts of abusive content online.
4. Measurement studies from academics and thinktanks indicate that 0.001% to 1% of content on mainstream platforms contains abuse. This is higher than the amount taken down the platforms but still suggests prevalence is low. However, some users and events generate far more abuse, such as prominent figures (e.g. MPs) and terror attacks.
5. Niche online forums (such as 4chan and Gab) can contain far more abuse than mainstream platforms, and in some cases between 5-8% of content is abusive or aggressive. These forums attract far fewer users and are not widely known by most people. Most research in this domain has focused on hate speech analysis and there is a lack of research into interpersonal abuse, such as harassment.

“The available evidence is fragmented, incomplete and inadequate for understanding the prevalence of online abuse.”

6. In strong contrast to all published figures and statistics, a large number of people report being exposed to online abuse. Based on survey data, including previously unseen analyses from the Oxford Internet Survey (OxIS)¹, we find that between 30-40% of people in the UK have seen online abuse. We also find that 10-20% of people in the UK have personally been targeted by abusive content online. Our analysis of OxIS shows that experiences of online abuse vary considerably across demographics:
 - a. Ethnicity: Black people and those of ‘Other’ ethnicities are far more likely to be targeted by, and exposed to, online abuse than White and Asian people. Differences in experiences of online abuse according to ethnicity are shown in Figure 1.
 - b. Age: Younger people are more likely to be targeted by, and exposed to, online abuse. They also spend more time online, which may partly explain this relationship.
 - c. Gender: Surprisingly, our analysis of OxIS did not identify a substantial difference according to gender. However, we advise caution as other survey data suggests that gender plays an important role in shaping peoples’ experiences of online abuse.
 - d. The OxIS dataset does not contain information about whether respondents identify as transsexual.
 - e. People with disabilities observe more online abuse than people without disabilities.
7. Overall, our analysis suggests that whilst the prevalence of online abuse is low, especially in terms of content which is illegal or contravenes platforms’ guidelines, a significant proportion of the population (approximately one-third) are exposed to it. This is deeply concerning.

¹ We are grateful to the authors of OxIS for giving us permission to use this data, and to Dr. Grant Blank for facilitating our access. Please see OxIS’s website for more information about the survey, <https://oxis.oii.ox.ac.uk>.

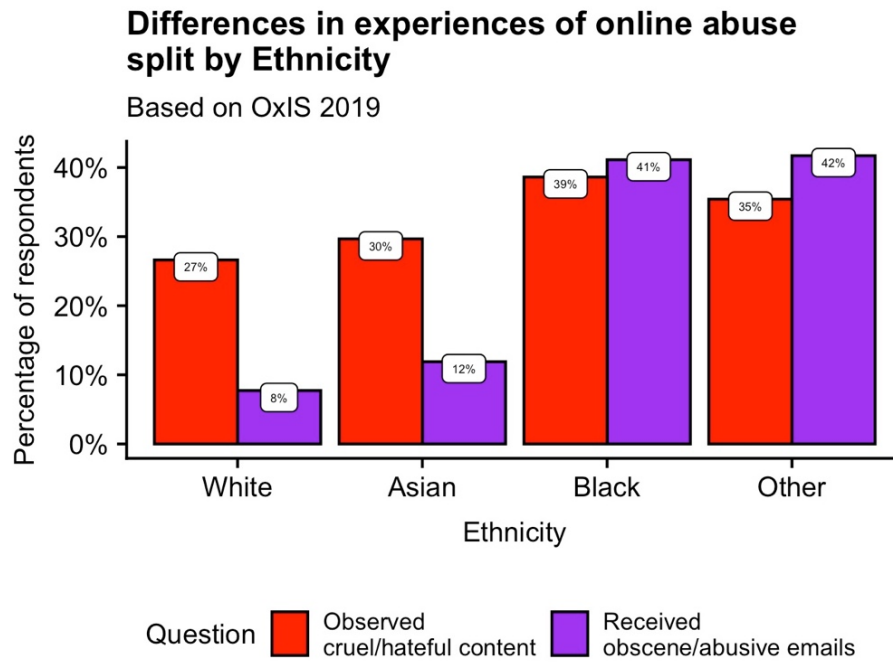


Figure 1: Experiences of online abuse, split by ethnicity (OxIS 2019 data)

Recommendations

Our review identifies some considerable shortfalls in existing monitoring practices for online abuse. We recommend:

1. A representative survey dedicated to understanding the experience of people in the UK of online abuse should be administered each year, rather than as a subsection of other surveys, such as OxIS and Ofcom's 'Adult media use and attitudes' survey. That said, both provide an excellent starting point.
2. Government statistics on different types of illegal online abuse, including both hate speech and online harassment, need to be centrally collated and published in a single bulletin. Efforts should be made to improve the coverage, comparability and quality of Government statistics, particularly re-instating online hate crime as part of the Home Office's reporting.
3. A publicly accessible monitoring platform should be established to provide real-time insight into the prevalence of online abuse. Whilst we recognise the limitations of computational tools, and of relying on 'big' rather than high quality datasets, efforts should be made to leverage recent computational advances, such as ensemble machine learning models, deep neural networks and contextual word embeddings.
4. Reporting standards for abusive online content need to be developed and backed by the Government. Many of the biggest platforms do not provide any information about online abuse and those which do only provide headline statistics – breakdowns by country and information about the targets and perpetrators of abuse are not given. Furthermore, platforms each use very different frameworks, guidelines, moderation processes and report content takedowns at different 'levels'. For instance, Twitter reports on the number of abusive users whilst Facebook reports on the number of abusive posts. Standardized reporting would ensure that the figures provided by platforms are directly comparable. We encourage Government to apply such reporting requirements to all large tech companies, including search platforms such as Yahoo, Google and Bing.
5. Researchers studying the prevalence of online abuse using observational methods, such as computational social scientific analyses, should explore more varied datasets and apply more nuanced detection tools. Noticeably, researchers have made little use of Google Trends data, which is freely available.²

² For one noticeable exception, see (Stephens-Davidowitz 2019).

“According to a 2019 survey by Ofcom, 18% of UK adults are concerned about hate speech on the internet (Ofcom 2019b).”

Background

Assessing the prevalence of online abuse is a difficult task. Even so, the UK public has expressed concern about its harmful effects: according to a 2019 survey by Ofcom, 18% of UK adults are concerned about hate speech on the internet (Ofcom 2019b). A survey of 500 women in the UK showed that 47% believe that current laws are inadequate (Amnesty International UK 2017a). A recent report from the Commission for Countering Extremism also found that 56% of the public believe ‘a lot more’ should be done to counter extremism online (Commission for Countering Extremism 2019). Many of the big tech companies have started responding to the public’s concerns, announcing plans to invest more money into content moderation, particularly the development of automated systems (YouTube 2019a). Civil society efforts to tackle online hate have also been stepped up through initiatives such as the recently founded Centre for Countering Digital Hate.³ As more resources are deployed to monitor, understand and counter hate it is crucial that we develop a better understanding of the scale of the problem (HM Government 2019).

Many examples demonstrate that the existing evidence base about online abuse is inadequate and that more comprehensive, detailed and accurate data is needed. In 2015, the European Commission on Race and Intolerance reported, ‘The actual extent to which hate speech is being used remains uncertain [...] This uncertainty is attributable to the absence of comprehensive and comparable data regarding complaints about the use of hate speech.’ (ECRI 2015, 20) Amnesty International similarly reported in 2017 that for hate speech, ‘reliable statistics are hard to come by.’ (Amnesty International UK 2017b) and the Home Offices’ 2018 thematic review of evidence on hate crime noted, ‘understanding the true prevalence of all forms of hate crime, particularly at a sub-strand level, remains a challenge’ and that knowledge of online offending is ‘incomplete’ and ‘patchy’ (Home Office 2018, 3, 5, 13).

³ The Centre for Countering Digital Hate’s website is available at <https://www.counterhate.co.uk>

Numerous policy reports also highlight the need for better evidence, including DCMS and the Home Office's Online Harms white paper, the Law Commission's review of hate speech laws (The Law Commission 2018), and the online harms 'rapid evidence assessment' from academic researchers supported by DCMS (Davidson et al. 2019). The implications of this lack of evidence for Government policymaking are severe. As Victoria Nash put it in her response to the Online Harms white paper, 'in the absence of a mature and rigorous evidence base, it is hard to see how [...] the broad ambitions to regulate [are justified].' (Nash 2019, 6).

Statistics on illegal online abuse provide one means of understanding the prevalence of online abuse. However, there are many examples which show that the relevant UK laws lack clarity and precision; it is not always clear which laws abusive content contravenes (The Law Commission 2018). During the Home Affairs Select Committee's evidence gathering on online hate crime in 2017, Carl Miller, research director of the thinktank *Demos*, described the 'confusion' around existing laws and noted, 'We have not had a proper law passed on this since social media became in widespread use.' (HM Government 2017, 18). In many cases, online flags are used inconsistently, which makes it hard to use official statistics representatively (Home Office 2018). A further constraint is that, in order to protect freedom of speech, the UK law sets a 'high bar' for what constitutes harassment and hate speech. These factors mean that statistics on illegal abuse do not provide a full picture of the online landscape.

Social media platforms have received considerable attention as one of the main spaces in which hate and harassment spreads online (Gorrell et al. 2018; Hine et al. 2017; Williams and Burnap 2016). Across Europe, Governments have responded to this by establishing reporting standards that online platforms must adhere to. Germany's Network Enforcement Act (NetzDG) requires platforms with more than 2 million users in Germany to publish semi-annual transparency reports if they receive more than 100 complaints per year, outlining: companies' content removal procedures, the number of complaints they receive and their source, the number of content takedowns (and the reasons why), and the amount of resources they dedicate to content moderation (Human Rights Watch 2018; Tworek and Leerssen 2019). In 2019 France passed a similar 'anti-hate' law to Germany, which establishes fines of up to 1.25 million Euros for companies which do not remove 'obviously hateful' content and requires them to report their moderation practices and resources (The Guardian 2019b). NetzDG and France's new anti-hate law make crucial information available for ensuring proper oversight by the public. However, NetzDG has been criticised for allowing platforms to 'come up with their own individual reporting formulas, making it difficult for users to flag NetzDG violations in a consistent and streamlined manner,' (Echikson and Knodt 2018, ii). In the UK, such reporting requirements are not mandated by law, although this is under consideration in DCMS and the Home Office's 'Online harms' white paper (HM Government 2019). Until platforms are given more direction in how to report on online abuse, it is likely that the information they provide will remain incomplete and partial.

Other sources of evidence can also provide insight into the prevalence of online abuse, such as focus groups, surveys, measurement studies and civil society monitoring groups. However, there are many missed opportunities in existing data collection processes using these methods. In 2018 the Equality and Human Rights Commission conducted ‘the first national survey of prejudice in Britain for over a decade.’ However, this survey did not provide any insight into online abuse as it ‘did not ask separately about online experiences’ (Abrams, Swift, and Houston 2018, 27). Similarly, an otherwise comprehensive report by the European Agency for Fundamental Rights on hate crime data collection practices across Europe did not contain any specific discussion of online content (FRA 2018) and the OSCE’s monitoring of online hate makes no provisions for whether offences are committed online (OSCE ODIHR 2018). In the long-term, monitoring of online abuse needs to be more closely integrated into existing monitoring frameworks to improve our understanding of the scale of the problem. In particular, although it is difficult because online abuse operates globally and often overruns national borders, we need more UK-specific estimates.

Defining abuse

A challenge for all researchers in the field of online abuse is how it should be defined and what actually constitutes abuse (Brown 2017). Part of the problem is that defining online abuse is not a purely scholastic endeavour: just by labelling content as ‘abusive’ we potentially create a moral and social impetus to constrain and challenge its dissemination online (Keck 2016; Modood et al. 2006; Simpson 2018). Once applied to real-world research tasks, definitions of abuse fast become embroiled in contentious debates around privacy, freedom of speech, democracy, discrimination and the power of big tech companies. To account for this, we adopt ‘minimal’ definitions. However, we caution that they are not necessarily applicable for all contexts and types of research, and that all definitions have limitations.

Existing research usually separates online abuse into two main varieties: abuse directed against a group, which is usually called ‘hate speech’ or ‘intergroup prejudice’, and abuse directed against an individual, which is usually called ‘harassment’ or ‘cyberbullying’ (Vidgen et al. 2019; Waseem et al. 2017). Abuse directed against individuals includes statements such as ‘I hate you’ or ‘@USERNAME you tw*t’ and group-directed abuse includes statements such as ‘Anyone wearing a hijab better watch out or else I’m going to stab them’ or ‘I want to kill all rapugees’. This distinction is embodied in the UK’s legal system, which has separate laws for online hate and online harassment (SELMA 2019; Williams and Pearson 2016). However, whilst this analytical distinction is useful, we caution that, in practice group- and person- directed abuse can be difficult to separate. For instance, female politicians can receive more abuse than male politicians. Some argue that although much of the abuse targets their individual traits and not their identity, they only receive personal abuse *because* of their identity. In such situations, it is ambiguous as to whether the abuse should be categorised as group- or person- directed.

A further consideration is how *strong* content must be for it to be considered abusive. UK Law and the big tech platforms' policies only protect against explicit hostility, such as making threats or using 'dehumanizing' language. In contrast, academic studies of prejudice have drawn attention to more 'subtle' forms, such as microaggressions and 'everyday' acts (Bliuc et al. 2018; Nadal et al. 2012; Pettigrew and Meertens 1995). Whether subtle forms of abuse should be monitored, and as such potentially moderated, is a highly contentious issue. In many cases, what some individuals consider to be 'subtle' forms of abuse, others might consider to be 'legitimate' forms of critique and everyday language (Salminen et al. 2018). Although there is a pressing need to support victims of abuse, protect potential victims and ensure public discourse is not toxified, caution is needed when developing policy interventions (Cohen-Almagor 2011).

Finally, many interventions in the social sciences suggest that whether content is considered abusive depends on both (i) individuals' background, experiences and attitudes (people with different outlooks will have very different opinions about whether content is abusive) (Butler 1997; Matsuda et al. 1993) and (ii) the context in which it is produced; work by Susan Benesch suggests that who speaks, when, with what authority, to whom and with what intention affect whether speech is genuinely hateful or, in some cases, 'dangerous' (Benesch 2012). As such, studies which purport to be studying 'abuse' may be studying different aspects of it or the same aspect in very different ways. In particular, respondents to surveys may have highly divergent opinions about what constitutes abuse and as such interpret questions differently. These factors should all be considered when evaluating the evidence presented here as the concept of 'abuse' is complex in any setting.

For the purposes of this report, we use two definitions:

1. In accordance with the European Commission against Racism and Intolerance, we define group-directed abuse as: 'the advocacy, promotion or incitement of the denigration, hatred or vilification of a person or group of persons, as well any harassment, insult, negative stereotyping, stigmatization or threat of such person or persons and any justification of all these forms of expression.' (ECRI 2015, 16).
2. Due to the lack of available definitions for person-directed attacks, we adapt the definition of cyberbullying offered by Dadvar et al. (Dadvar et al. 2013): aggressive, intentional acts carried out by a group or individual, using electronic forms of contact against an individual.

The prevalence of online abuse in the UK

1. UK Government statistics on illegal online abuse

1.1 Key findings from UK Government statistics

- 1.1.1 1,605 online hate crimes were recorded in England and Wales in 2017/18 (Home Office 2018). This information has not been made available by the Home Office for 2018/19.
- 1.1.2 We estimate that the per capita rate of offences under Section 127 of the Communications Act 2003 is less than 200 per 100,000 people and the rate of arrests is ~10 per 100,000 people.
- 1.1.3 The Scottish Crime and Justice Survey found that around ~2% of respondents had experienced online abuse (Scottish Government 2018). Comparable results are not available for England, Wales and Northern Ireland.
- 1.1.4 To improve the quality of reporting and to provide a complete picture in the future, we present a 'wishlist' for Government partners:
 - 1. To ensure a UK-wide evidence base, the same figures should be published for Scotland and Northern Ireland as England and Wales. Each nation's respective crime surveys should also contain similar questions.
 - 2. Statistics for how much criminal online abuse is directed against individuals should be collated and presented in a single report, covering offences under all of the relevant laws.
 - 3. Statistics for the number of offences, arrests and prosecutions, as well as offenders, victims, arrestees and people prosecuted, should all be provided in one report. This will ensure that data is joined-up and can be interpreted easily.
 - 4. Provided that it does not risk making personally identifiable information available (if, for instance, the sample sizes are very small) more detailed information should be provided, including the demographics of the offenders and victims, the geographic location of crimes and when they took place.
 - 5. Reporting online hate crime should be mandatory for all police forces.

“In strong contrast to all published figures and statistics, a large number of people report being exposed to online abuse.”

1.2 Background to the Law on Abuse

- 1.2.1 Laws prohibiting hate crime include the Public Order Act 1986, Crime and Disorder Act 1998 and Criminal Justice Act 2003. At present, only five aspects of identity are protected by UK law: disability, transgender status, race, religion and sexual orientation. Noticeably, four of the other nine ‘protected characteristics’ in the 2010 Equality Act are not protected: gender reassignment, marriage and civil partnership, pregnancy and maternity, and sex (Equality Act 2010, Equality and Human Rights Commission). In 2018, there were repeated calls for misogyny to be made a hate crime, but this was resisted by, amongst others, UK Police forces (BBC News 2018). This impacts the recorded prevalence of abuse: if the law protected a wider array of identities from abuse then the recorded prevalence would be higher. Under-reporting may also affect some types of illegal behaviour more than others. Amnesty International report that one in two racist hate crimes are reported to the police, one in four homophobic hate crimes, one in 10 religiously motivated hate crimes, and one in 19 disability hate crimes (Amnesty International UK 2017b).
- 1.2.2 Online harassment, which we describe as a type of ‘person-directed’ abuse, is primarily prohibited through the Malicious Communication Act 1998 and Communications Act 2003. The law is clearer in this area than with hate speech, with Section 127 of the Communications Act 2003 providing online-specific provisions against person-directed abuse. However, critics have still raised concerns about the narrow scope of the law and how it is enforced (Davidson et al. 2019; The Law Commission 2018; Williams and Pearson 2016).
- 1.2.3 The criteria determining whether content can be considered illegal is intentionally set stringently so as to avoid constraining freedom of speech (The Law Commission 2018). As such, the statistics report here pertain only to the most extreme forms of online abuse.

1.3 Measuring online hate crime (group-directed abuse)

1.3.1 The Home Office's 2017/18 report on hate crime in England and Wales shows that during the year there were 94,098 hate crimes, both online and offline (Home Office 2018). From 2015 onwards, police forces have been required to add an 'online flag' to reports of crimes, and this has been published in the Home Office's reports since 2016/17.

However, in the 2017/18 report, only 30 out of 44 police forces provided data of sufficient quality for this to be included in statistics and this is classified as an 'experimental' result because it does 'not meet the rigorous quality standards of Official statistics.' In particular, due to inconsistencies in how the 'online flag' for reporting online hate crime is used, the Home Office acknowledges that, 'any figures produced using the different flags are likely to be underestimates.' (Home Office 2017, 18).

1.3.2 1,605 hate crimes had an online element (2.3% of the total). Of these, 80% were for 'Violence against the person', 14% for 'Public order offences', 6% for 'Other notifiable offences' and 0.2% 'Criminal damage and arson'. Because statistics for online hate and online harassment are compiled separately, it is unclear how many of these offences for online hate are also considered harassment or whether they fall under other laws.

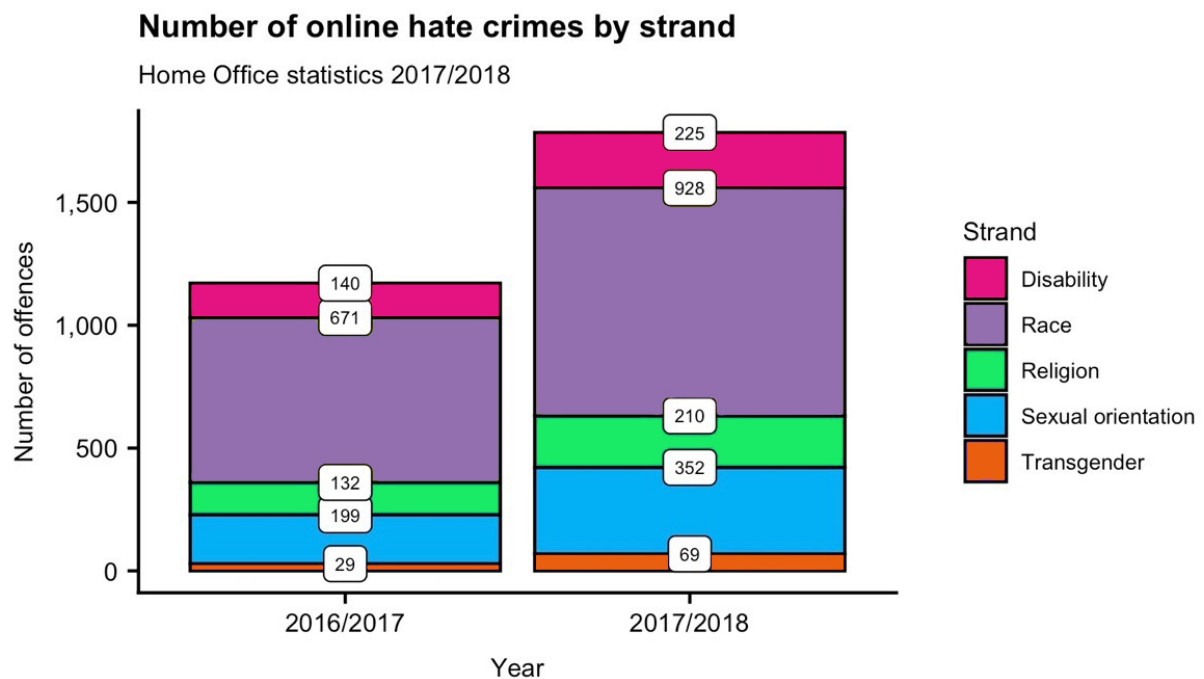


Figure 2: Online hate crime by strand, 2016/2017 to 2017/2018

- 1.3.3 The breakdown of online hate crimes by target in England and Wales is shown in Figure 2. Note that statistics for online hate crime are not currently available for Northern Ireland or Scotland (Northern Ireland Statistics and Research Agency 2019; Scotland Prosecution Service 2017).

1.4 Measuring harassment (person-directed abuse)

- 1.4.1 Assessing the prevalence of illegal abuse directed against individuals online faces similar challenges to assessing the prevalence of online hate: figures are not collated in a single location and online specific figures are not usually published. At present, Freedom of Information requests provide the most valuable information. Section 127 of the Communications Act is the most directly relevant part of UK Law, and we use this as our focus.
- 1.4.2 The Metropolitan Police is responsible for 8.8 million people (Metropolitan Police 2017). From 2006/07 to 2015/16, the per capita rate of offences under Section 127 was less than 200 per 100,000 people and the rate of arrests was less than 12 per 100,000 (as 1 in 20 offences for online harassment led to an arrest (~4.3%)) (The Metropolitan Police 2016).
- 1.4.3 In 2017, the police forces of the West Midlands, Hertfordshire and Merseyside made 569 arrests. They are responsible for policing 5.5 million people in total (Hertfordshire Constabulary 2018; Merseyside Police 2019; West Midlands Police 2019). As such, for all three Police Forces, the per capita rate of arrests was 10 per 100,000.
- 1.4.4 The Scottish Crime and Justice Survey asks respondents about experiences of being insulted or harassed online. For 2017/18 it reported that 14% of adults had been insulted or harassed outside of their homes and that, of these, 16% had encountered such behaviour digitally (Scottish Government 2018, 102). This equates to around ~2% of respondents. 58% of these victims of online abuse were female and 42% were male (Ibid., p. 125).

“The implications of the lack of evidence for Government policy-making are severe.”

2. Civil Society reports of online abuse

2.1 Results

- 2.1.1 Civil society groups play an important role in monitoring online abuse by receiving and checking reports from users. Police forces currently work with several partners via the 'True Vision' website.⁴ Users report incidents to the partner organisations and then trained experts work with them to verify that incidents took place in the UK and that they genuinely constitute hate. Incidents are reported to the police if they are likely to be illegal. We draw on data from two established partners: The Community Security Trust (CST), which campaigns for the security and safety of Jewish people in the UK, and Tell Mama, which provides support to victims of Islamophobia.
- 2.1.2 Tell Mama reported 362 verified online Islamophobic incidents in 2017 (30% of all the reports they received), 311 in 2016 (33%) and 364 in 2015 (45%) (Tell Mama 2018).
- 2.1.3 The CST reported 384 verified online anti-Semitic incidents in 2018 (23% of all the reports they received), 249 incidents in 2017 (18%), 289 in 2016 (21%) and 185 in 2015 (19%) (CST 2019).
- 2.1.4 Tell Mama and the CST each report more religious based hate incidents than the total number of illegal online religious hate crimes in each year: across all religions, 210 online hate crimes were flagged in 2017/18 and 132 in 2016/17. However, the total number of incidents recorded by both organisations is still relatively low, and far less than 1,000 for every year.
- 2.1.5 There are three challenges with using user reports as a source of evidence into the prevalence of online abuse. First, the reporting process is detailed but also onerous and this means that under-reporting is like to be high. Second, results may be affected by changes in reporting, such as the publicity efforts of the partner organisations, which means that longitudinal comparisons cannot be made. Third, the reporting processes set a high threshold for what constitutes hate and as such will miss many less overt instances. Thus, whilst these figures are useful for building up a picture of which groups are most targeted by hate, they only provide a partial and narrow view.

⁴ For further information about the 'True vision' website see: <http://www.report-it.org.uk/home>

	CST (anti-Semitism monitoring)	Tell Mama (Islamophobia monitoring)
Number of online incidents (2015)	185	364
Number of online incidents (2016)	289	311
Number of online incidents (2017)	249	362
Number of online incidents (2018)	384	N/A

Table 1: Number of online incidents reported to CST and Tell MAMA, 2015 to 2018.

3. Platforms' transparency reports on abuse

3.1 Key findings from platforms' reports on abuse

- 3.1.1 The prevalence of online abuse which meets the stringent content moderation requirements set by platforms is very low (~0.001% of all content hosted). Four of the major platforms provide sufficiently detailed transparency reports to estimate the proportion of hosted content which is actioned for being hateful or harassing. We suggest it is around 0.001%, although this varies considerably by platform and our estimate is based on third-party sources and heuristics.
1. 0.001% of content posted on Facebook.
 2. 0.001% of videos posted on YouTube.
 3. 0.0001% of content posted on Reddit.
 4. 0.2–0.3% of users on Twitter.
- 3.1.2 The major platforms do not segment figures on online abuse by which groups are targeted by abuse, who sends the abuse, and what country it takes place in. This limits our ability to understand at a granular level how much abusive content they host. We welcome efforts by the UK Government to provide further guidance in this area and to provide a regulatory framework for social media platforms to report online abuse.
- 3.1.3 Criticisms have been raised against the content removal policies and practices of platforms, with concerns raised about whether are insufficiently proactive and leave substantial amounts of abuse online. As such, it is possible that the figures which are reported here do not reflect all of the abuse that is hosted by the big platforms.
- 3.1.4 In this report, we focus on understanding the amount of content, users, groups and channels which are abusive. This is a useful starting point for characterizing the prevalence of online abuse – but, arguably, it would be more insightful to understand the amount of times that online abuse is viewed or interacted with. This better characterizes the role of online abuse within the digital environment because it provides more context about how much users are actually exposed to it. Measuring the number of views/impressions of content is more demanding and, to our knowledge, Facebook is the only major platform which has sought to make such information available. The technical and social challenges in reporting such information needs to be explored further across all platforms.

- 3.1.5 We provide the following ‘wish list’ for platforms to enhance their future reporting:
1. Provide separate figures for each country.
 2. Provide separate figures for the amount of content, users and ‘spaces’ (i.e. groups, pages or channels, depending on the platform) which are actioned.
 3. Provide breakdowns for the targets of abuse (e.g. gender, ethnicity, age).
 4. Provide estimates for the total amount of content, users and ‘space’ that is hosted so figures for abusive content can be meaningfully interpreted.
 5. Stipulate in reports when a change in policy has been implemented and provide an estimate of how much additional ‘old’ content this has affected by the new policy.

“A publicly accessible monitoring platform should be established to provide real-time insight into the prevalence of online abuse.”

3.2 Overview of platforms’ reports

- 3.2.1 Transparency reporting by social media platforms is a useful source of data on online abuse. However, it is not mandated by UK Law and as such many platforms do not provide any information. Pew Research reports that the most widely used social media platforms in 2019 were Facebook, Twitter, YouTube, Snapchat, Instagram, WhatsApp, LinkedIn, Pinterest and Reddit (Pew Research 2019). Of these nine platforms, only four, Facebook, Twitter, YouTube and Reddit, provide statistics on online abuse as part of their transparency reporting. Many platforms which are not ‘social media’ companies but are nonetheless major digital content providers, including search providers, such as Google and Bing, and news platforms, such as BBC News, do not provide transparency reports of any kind.
- 3.2.2 None of the platforms surveyed provide UK-specific figures. As such, we use their figures for global harmful content monitoring and must advise caution when applying them to the UK. Furthermore, platforms’ reporting standards vary considerably, and each provides different and non-comparable information.

- 3.2.3 The most recent transparency reports for all four platforms we have studied show they are now actioning more content than before, through bans, suspensions and warnings. However, it is unclear what has driven this increase. It could be that (i) the amount of abuse has increased, (ii) platforms are enforcing more stringent policies or (iii) the technology used by platforms is flawed and there are a lot of false positives (i.e. content which is not abusive but is erroneously labelled as such is being taken down). We speculate that it is a mix all three and as such advise caution when interpreting increases in the amount of abusive content taken down by platforms as it may not indicate an increase in the amount of abusive content *per se*.
- 3.2.4 Twitter and Facebook apply newly updated policies and content moderation technologies to *all* of the content they host.⁵ We have made the same assumption for Reddit and YouTube, as the total amount of content they each host is unknown.
- 3.2.5 We provide estimates for the prevalence of abusive content. However, we cannot estimate how many users have been exposed to such content. If platforms quickly remove abusive content after it has been uploaded (or stop it from being uploaded straight away) then the percentage of 'impressions' for abusive content could be far lower than the percentage of posts which are abusive.
- 3.2.6 Figure 3 shows the percentage change in the amount of abuse actioned each period, compared with the first reported period. Figures are reported for YouTube, Facebook and Twitter. Reddit is not shown as they have only completed one period of transparency reporting. The results show (i) the considerable gaps in current reporting and the lack of longitudinal data, (ii) the marked increase in the amount of hateful content removed by YouTube in Q2 2019 and (iii) the small decrease in the amount of content removed by Twitter within both the Bullying and Hate category.

3.3 Facebook

- 3.3.1 Facebook currently measures abuse by how much content is actioned (Facebook 2019). Figures for the number of abusive individuals or pages are not reported. An additional metric, 'prevalence' captures the frequency at which content that violates the Community Standards is viewed, and has been used to measure several types of online harm. However, for both 'hate speech' and 'bullying and harassment', it is not currently reported.

⁵ The policies of Reddit and YouTube suggest that they also apply newly updated policies and technologies to all of the content they host.

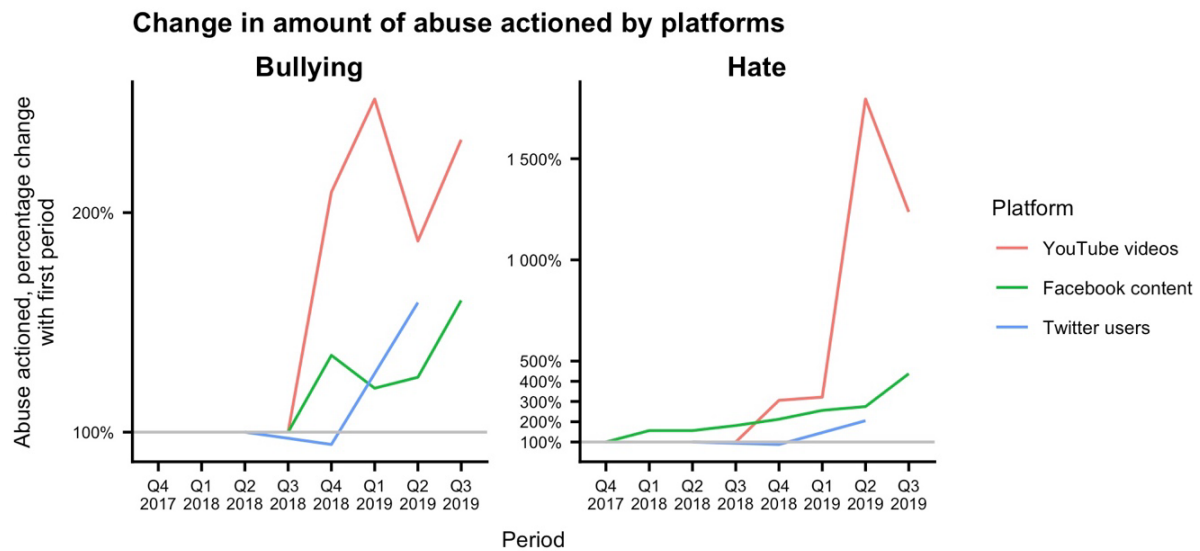


Figure 3: Change in the amount of abuse actioned by major platforms, 2017-19.

- 3.3.2 Statistics on online abuse were first published in Q4 2017, and the most recent report covered Q3 2019. The amount of actioned ‘hate speech’ content has increased quarter-on-quarter, from 1.6 million pieces in Q4 2017 to 7.0 million pieces in Q3 2019 (see Table 2). Less content is actioned for ‘bullying and harassment’, for which 3.2 million bits of content were actioned in Q3 2019.
- 3.3.3 Understanding what *proportion* of the content hosted on Facebook is abusive is difficult as the platform does not provide regular updates about the total amount of content it hosts. In 2013, Facebook reported that, globally, 4.75 billion posts are made each day (Robertson 2013). The accuracy of this figure in 2019 is debatable; although the number of active users on Facebook has increased, some suggest that users post less frequently (CBS News 2019). Furthermore, users may remove content, especially older posts, at a high rate which makes this figure even harder to estimate. As such, we can only estimate the prevalence of abusive content on Facebook using heuristics.
1. Taking 2013 as our starting point, we estimate that there has been a net increase of 0.5 billion posts per day up until 2019, accounting for changing posting levels and rates of content removal by users. In total, we estimate there are one trillion posts on the platform, although this figure is speculative. It suggests that from Q1-Q4 2018, the rate of content which was actioned for either ‘hate speech’ or ‘bullying and harassment’ was 0.001%.

Category	Q4 2017	Q1 2018	Q2 2018	Q3 2018	Q4 2018	Q1 2019	Q2 2019	Q3 2019
Amount of content actioned (hate speech)	1.6M	2.5M	2.5M	2.9M	3.4M	4.1M	4.4M	7.0M
Amount of content actioned (bullying and harassment)	N/A	N/A	N/A	2.0M	2.7M	2.4M	2.5M	3.2M

Table 2: Content actioned by Facebook for hate speech and bullying/harassment
(Community Standards Enforcement Report, May 2019, Facebook)

3.4 YouTube

- 3.4.1 YouTube first published statistics on online abuse in Q3 2017, when the number of videos *flagged* for hate were shared. In Q3 2018, transparency reporting was expanded to include the number of videos, channels and comments which are *removed* for 'hateful or abusive', as well as 'harassment and cyberbullying'. The most recent report covers Q3 2019 (YouTube 2019b).
- 3.4.2 From Q4 2018 to Q3 2019, 12.6 million channels were removed by YouTube. 46,197 were for 'harassment and cyberbullying' and 30,412 were for 'hateful or abusive'.
- 3.4.3 In the same period, 34.9 million videos were removed. 166,103 of these were for 'harassment and cyberbullying' and 226,676 for 'hateful or abusive'. In Q2 2019 there was a noticeable increase in the amount of videos removed for 'hateful or abusive', up to 111,185 videos. Nonetheless, it remains one of the less prevalent forms of online harm which the platform tackles.
- 3.4.4 In Q3 2019, YouTube also removed 517 million comments for being harmful, following a two-fold increase in the previous quarter. It is unknown how many of these comments were removed for being either hateful or bullying and how many comments were made in total during the period.

3.4.5 YouTube does not report the total amount of channels and videos it hosts. Third party sources suggest that it has at least over 23 million channels⁶ and that over half a million hours of video are uploaded every day (Tubics 2018). As such, we estimate prevalence using plausible heuristics:

1. If there were 23 million channels on YouTube during the year from Q4 2018 to Q3 2019 then 0.25% were removed for being 'hateful or abusive' or 'harassment and cyberbullying'. However, we caution that the figure of 23 million is most likely a considerable under-estimate, thereby inflating the prevalence of abusive channels.
2. If there were 100 million channels on YouTube during the year from Q4 2018 to Q3 2019 then 0.06% were removed for being 'hateful or abusive' or 'harassment and cyberbullying'.
3. If there are ten billion videos on YouTube then the rate of removal for 'hateful or abusive' content was 0.002% and for 'harassment and cyberbullying' 0.001% from Q4 2018 to Q3 2019.
4. If there are 100 billion videos on YouTube then the rate of removal for 'hateful or abusive' content was 0.0002% and for 'harassment and cyberbullying' was 0.0001% from Q4 2018 to Q3 2019.

⁶ Social Blade reports that this figure does not take into account channels with fewer than 5 subscribers and has uneven coverage of those with between 5 to 10 subscribers, which may account for a large number of the total (Tubics 2018).

	Category	Q3 2018 ⁷	Q4 2018	Q1 2019	Q2 2019	Q3 2019
Channels (removed)	Total number of channels removed	1.7m	2.4m	2.8m	4.1m	3.3m
	Harassment and cyber-bullying (% of total channels removed)	1,757 (0.3%)	8,061 (0.3%)	10,623 (0.4%)	14,668 (0.4%)	12,845 (0.4%)
	Hateful or abusive (% of total channels removed)	524 (0.1%)	1,713 (0.1%)	3,379 (0.1%)	17,818 (0.4%)	7,502 (0.2%)
Videos (removed)	Total number of videos removed	7.8m	8.8m	8.3m	9m	8.8m
	Harassment and cyber-bullying (% of total videos removed)	18,848 (0.7%)	39,456 (0.5%)	47,443 (0.6%)	35,257 (0.4%)	43,947 (0.5%)
	Hateful or abusive (% of total videos removed)	6,195 (0.2%)	18,950 (0.2%)	19,927 (0.2%)	111,185 (1.2%)	76,614 (0.9%)
Videos (flagged)	Total number of flags given to videos ⁸	42m	44.5m	48.5m	45m	49m
	Hateful or abusive (% of total flags given)	7,432,077 (17.7%)	7,562,626 (17.0%)	8,187,076 (16.9%)	7,750,866 (17.2%)	9,705,685 (19.8%)
Comments (removed)	Total number of comments removed	166m	190m	278m	538m	517m

Table 3: Channels, videos and comments flagged/removed by YouTube for hate speech and cyberbullying/harassment.

⁷ For Q3 2018, figures for channels and videos (removed) only cover September.

⁸ Each video can be flagged multiple times so the total number of flagged videos is lower than the number of flags given. YouTube does not provide a breakdown by type of content violation for comments.

3.5 Reddit

- 3.5.1 Reddit reports on abusive content takedowns at one level: pieces of content. Reddit published its first transparency report in 2014. However, statistics on abusive content were only first included in 2018, which is also the most recent year for which information is available.
- 3.5.2 In 2018, Reddit received 221,116 content policy violation reports. Of these reports, 79,602 (36%) were actioned. 14,806 (18.6%) were for 'harassment' and 12,657 (15.9%) were for 'encouraging violence', which includes self-harm (Transparency Report 2018). Reddit does not include 'hate speech' as a category in its transparency reporting. In 2018, in addition to platform-level policy violations, 50,547,715 pieces of content were removed by community moderators. No information is available as to what percentage hate speech and harassment comprise.
- 3.5.3 As of 2018, Reddit claims that it has over three hundred and thirty million monthly active users and over 130,000 active communities, which often comprise more than just one subreddit (Reddit 2019). It does not provide figures for the total amount of content it hosts. Third party sources estimate that around 3 million comments are made each day, although this figure has not been confirmed by the company (DMR 2019).
1. If there are 10 billion bits of content on Reddit then 0.00027% were removed in 2018 by the platform for 'harassment' or 'encouraging violence'.
 2. If there are 100 billion bits of content on Reddit then 0.000027% were removed in 2018 by the platform for 'harassment' or 'encouraging violence'.
 3. Even if community moderators only removed hate speech and bullying, which is entirely unrealistic, then 0.5% of content would fall under these categories, based on the platform hosting 10 billion bits of content.

3.6 Twitter

- 3.6.1 Twitter reports on abusive content at one level: accounts (Twitter 2019b). Noticeably, it does not report on the prevalence of abusive tweets. Twitter published its first transparency report in 2012. Statistics on abusive content were only included in 2018 and the most recent report covers Q1/2 2019. Twitter uses three categories for moderating online abuse: 'abuse' (which includes harassment and intimidation), 'hateful conduct' and 'violent threats'.
- 3.6.2 During Q1-Q4 2018, 6.5 million accounts were reported for 'abuse'. Of these, ~480,000 (7%) were actioned.

- 3.6.3 During Q1-Q4 2018, 6.2 million accounts were reported for 'hateful conduct'. Of these, ~540,000 (8%) were actioned.
- 3.6.4 During Q1-Q4 2018, 3 million accounts were reported for 'violent threats'. Of these, ~100,000 (3%) were actioned.
- 3.6.5 Twitter reported that it has 330 million active users each month as of Q1 2019 (Twitter 2019a). We use this as a heuristic to estimate prevalence:
1. If there were 330 million users on Twitter during Q1-Q4 2018 then 0.3% were actioned for any type of online abuse (1,119,811 in total).
 2. However, the figure of 330 million users is likely an underestimate as many accounts which were actioned may have been either inactive or were quickly recreated after being taken down. Thus, the relatively stable figure of 330 million could be hiding considerable 'churn' in the number of accounts which are created or taken down. If there were 500 million accounts created in total during Q1-Q4 2019, then 0.22% were actioned for any type of online abuse.
- 3.6.6 Prior to publishing statistics on abusive content in its transparency reports, Twitter released three 'Government Terms of Service' reports, which cover Q3-Q4 2016, Q1-Q2 2017 and Q3-Q4 2017.⁹ In Q3-Q4, 10,323 pieces of content were reported for abusive behaviour, of which 12% were for hateful content and 16% were for harassment. 6,885 accounts were also reported during this period, of which 6,254 were for abuse (91%). 25% of these 6,254 accounts were subsequently actioned.

⁹ (Abusive Behaviour reports, Government TOS reports – January to June 2017, Twitter Rules enforcement, Transparency Report, Twitter).

Accounts	Q1/Q2 2018	Q3/Q4 2018	Q1/Q2 2019
Number of accounts reported (abuse)	2.8m	3.7m	4.6m
Number of accounts actioned (abuse)	248,629 (9%)	235,455 (6%)	395,917 (9%)
Number of accounts reported (hateful conduct)	2.7m	3.5m	5.2m
Number of accounts actioned (hateful conduct)	285,393 (11%)	250,806 (7%)	584,429 (11%)
Number of accounts reported (violent threats)	1.4m	1.7m	2.0m
Number of accounts actioned (violent threats)	42,951 (3%)	56,577 (3%)	56,219 (3%)

Table 4: Accounts reported and actioned by Twitter for abuse, hateful conduct and violent threats.

4. Measurement studies

4.1 Key findings from measurement studies

- 4.1.1 Measurement studies show that the prevalence of abuse is less than 1% on mainstream platforms.
- 4.1.2 However, some users and events generate more abuse: studies show that 1-2% of tweets associated with contentious political events and 2-5% of tweets targeted at MPs are abusive.
- 4.1.3 There is a lack of publicly available evidence into many online spaces, including search data (such as Google and Bing), some mainstream platforms such as Instagram and Facebook, and the comments sections of news websites.
- 4.1.4 Niche alternative spaces contain far more abuse, and in some spaces between 5-8% of content is abusive. However, most research in this domain has focused on hate speech analysis and there is a lack of research into person-directed abuse, such as harassment.

4.2 Overview

- 4.2.1 Data-driven studies by academics, civil society groups and think tanks are a useful way of supplementing other sources of evidence to build a broader and more detailed picture of the prevalence of online abuse. They can be split into two types: (i) those which investigate abuse within mainstream online spaces, such as Facebook and Twitter and (ii) those which investigate abuse within niche and marginal spaces, such as Gab and alternative communities on Reddit.
- 4.2.2 A key advantage of measurement studies is that they tend to provide greater insight into both the prevalence of hate and its dynamics, including network, temporal and causal analyses (Bakalis 2018; Bliuc et al. 2018). However, measurement studies often rely on relatively small amounts of data and, as Davidson et al. noted in a recent assessment of available evidence on online harms, ‘cannot do justice to the scale of a problem which, internationally, is vast[.]’ (Davidson et al. 2019)
- 4.2.3 We identify three specific problems which should be considered when interpreting the results of measurement studies:
 - 1. Studies are often unrepresentative and very specific, focusing on only certain platforms, communities, time periods, types of users and contexts. Many focus on just Twitter and do not consider other online platforms, which can introduce considerable biases in terms of the demographics of users (Blank 2017; Cihon and Yasserli 2016).

2. In many cases studies are either focused on non-UK geographic areas or do not explicitly analyse users' location. Geolocating online users remains a considerable research challenge, beset with conceptual, methodological and ethical difficulties. Nonetheless, this still limits the direct relevance of many studies to understanding the UK context.
3. Tools for measuring online abuse are still only nascently developed and often have a high level of error when applied in 'the wild.' These limitations mean that results should be treated with caution (Vidgen et al. 2019).

4.3 Mainstream platforms report low levels of abuse overall

- 4.3.1 Measurement studies, in which abusive content detection systems are applied in real world settings, offer alternative insight into the prevalence of online abuse. We summarise findings from several recent measurement studies on mainstream platforms.
- 4.3.2 In order to create abusive content detection systems, many studies in computer science involve researchers manually labelling large amounts of data, such as social media posts. This highly laborious task is a fundamental part of creating a *machine learning* system: it provides the data which the machine trains on.

Such research can also be used to provide estimates of the amount of abusive content that is observed in 'the wild' (Schmidt and Wiegand 2017; Vidgen et al. 2019). Founta et al. developed a large dataset of annotated tweets for hate speech detection. They report that, depending on the categorisation schema which is used, the prevalence of abusive tweets is typically between 0.1% and 3% (Founta et al. 2018, 3). Wulczyn et al. similarly report that the prevalence of personal attacks on Wikipedia talk pages is 0.9% (Wulczyn, Thain, and Dixon 2017, 3) and Golbeck et al. also state that random sampling of tweets resulted in only a few dozen offensive ones being found (Golbeck et al. 2017). In a review paper, Wiegand et al. confirm this finding, noting that 'pure random sampling [...] would always result in tiny proportions of abusive content.' (Wiegand, Ruppenhofer, and Kleinbauer 2019, 603). Thus, based on a large number of independent studies, we suggest that the prevalence of any type of abuse is less than 1% when randomly sampled.

- 4.3.3 Researchers at the think tank Demos investigated the prevalence of Islamophobia on Twitter from March to July in 2016 (Miller et al. 2016). They used a machine learning algorithm, leveraged from a dataset of manually annotated tweets, to identify 657,650 Islamophobic tweets during the five-month period. These were out of a 34 million dataset which was collected from the Twitter API using a list of potentially anti-Islamic slurs (such as, 'Jihad', 'BNP' and 'kaffir'). All of the collected tweets were sent in English from geographic locations around the world. Approximately 2% of the identified tweets were Islamophobic.

A further study by Demos examined anti-Islamic tweets from the UK sent during March 2016 to March 2017 (Demos 2017). They identified 143,920 tweets, sent by 47,000 users which were geolocated in the UK. Another Demos study of Brexit identified 5,484 tweets sent from the UK between 19 June and 1 July 2016 which contained xenophobic or anti-immigrant attitudes (Miller, Krasodonski-Jones, & Dale, 2016). In both of these latter two studies, tweets were taken from the 1% Twitter stream and it is unclear what proportion of the data they comprise. Burnap and Williams study 210,807 tweets posted after the 2015 Woolwich terrorist attack and classify 1,878 (~1%) as containing BME or religious hate-related content (Williams and Burnap 2016). Overall, these results suggest that even tweets associated with contentious political events contain around 1-2% abuse. Note that each of these studies only examined specific types of abuse, and figures may be higher if all types of abuse were included in models.

- 4.3.4 Several studies have investigated the prevalence of abuse directed against prominent figures, such as MPs. From a dataset of 11.4 million sent against MPs from 9 May to 18 August 2016 (during the Brexit referendum), Krasodonski-Jones identified 188,000 abusive tweets (5%), using a machine learning classifier (Krasodonski-Jones 2017). Gorrell et al. collect a dataset of tweets directed against MPs for a month in 2015 and 2017, and identify 2.8% and 4% of tweets, respectively, as abusive, using a keyword search (Gorrell et al. 2018).

Subsequent work, which also includes results for a month in each of 2018 and 2019, shows similar prevalence of abuse (Greenwood et al. 2019). A report by Amnesty International, which examined a dataset of 900,223 tweets directed to the 177 women MPs active on Twitter from 1 January to 8 June 2017, found that approximately 26,000 were abusive (~3%) (Amnesty International UK 2017c). These studies differ in methodology and data collection, but suggest that 2-5% of the tweets received by MPs are abusive.

- 4.3.5 Concerns have been raised that the comment sections of news sites are rife with trolling, incivility, harassment and hate. However, statistics on the number of offensive posts are hard to find, with many platforms refusing to share even headline statistics. The Guardian reports that 1.4 million comments from 1999 to 2016 (2% of the 70 million total) were blocked, often because they were abusive. (The Guardian 2016). Cheng et al. study comments on CNN.com from December 2012 to August 2013 and report that 20,197 out of 865,248 users were banned (2.3%) and that 3.8 million out of 16.5 million posts were banned (23%). However, they do not report the reasons for bans (which can include spamming and posting non-relevant content as well as abuse).

- 4.3.6 Stephens-Davidowitz, in a study commissioned by the anti-Semitism support and monitoring charity CST, reports that 170,000 Google searches in the UK each year (2005 to 2018) are anti-Semitic, of which 10 percent involve violent language (Stephens-Davidowitz 2019). This number cannot easily be converted to a percentage, as billions of Google searches are made each day.

4.4 Niche platforms report higher levels of abuse

- 4.4.1 Research into niche online spaces provides higher estimates of the prevalence of online abuse, even suggesting that some communities, and entire platforms, are deeply hateful. These studies show the uneven way in which abuse manifests online.
- 4.4.2 Chandrasekharan et al. propose that certain subreddits, such as r/fatpeoplehate and r/CoonTown are inherently 'toxic' (Chandrasekharan et al. 2017). Similarly, research suggests that the white supremacy forum Stormfront is a hub of hateful activity, as are other sites such as the Daily Stormer (Bowman-Grieve 2009; Figea, Kaati, and Scrivens 2016). However, such claims may be overstated. De Gibert et al. qualitatively investigate a sample of posts from Stormfront and identify 1,119 out of 10,000 posts as being hateful (11%). Whilst this is far higher than in mainstream platforms, it suggests that many posts there are non-hateful. Furthermore, it is difficult to generalise from this figure and situate it in the broader social context: the number of Stormfront members is estimated at around 300,000, making over 1,000 posts per day—but this figure has not been confirmed and users' location is unknown (Figea, Kaati, and Scrivens 2016).
- 4.4.3 Gab has received considerable attention from academic researchers, since it was started in early 2016 as a 'free speech platform,' and quickly became used by the alt-right. Its Terms of Use are far more lax than for mainstream platforms, and do not explicitly prohibit hate speech or bullying, and only require that users are not 'unlawful' and do not 'unlawfully threaten' other users (Content Standards, Website Terms of Use, last modified 24 June 2019, Gab). Zannettou et al. investigate a dataset of 22 million posts from August 2016 to January 2018, posted by 336,000 users. They report that 5.4% of posts include hateful terms, identified using a lexicon of hateful terms from *Hatebase* and conclude that Gab is positioned at the 'border of mainstream social networks like Twitter and 'fringe' web communications like 4chan.' (Zannettou et al. 2018) Mathew et al. also investigate the prevalence of hate on Gab (Mathew et al. 2018). They use a lexicon of hateful terms, primarily comprising slurs, to identify hateful content. They find that out of a 21 million post dataset, collected from October 2016 to June 2018, just 167,782 posts (0.79%) are hateful. However, they note that their method is keyword-based and so misses more subtle forms of hate and also posts which are just images, videos and URLs.

The discrepancy between the figures reported by Zannettou et al. and Mathew et al. (almost one order of magnitude, from 5.4% to 0.79%) highlight the challenges of measuring online hate.

- 4.4.4 4chan and 8chan have both received considerable attention for their use by hateful online actors, including being the venue of several extremist manifestos in the aftermath of hate-motivated terror attacks (The Guardian 2019a). Hine et al. investigate a dataset of 8 million posts collected over 2.5 months during 2016 (Hine et al. 2017). They use a HateBase lexicon to measure the prevalence of hate and report results for 3 popular forums (or 'boards' on 4chan), showing: 12% of posts in the forum '/pol/' are hateful, 6.3% of posts in '/sp/' and 7.3% of posts in '/int/'. These results are then compared with a random sample of tweets from Twitter, in which just 2% of posts are identified as being hateful. This indicates that, whilst not all posts are hateful even in the more extreme parts of the internet, the level of hate is significantly higher than in mainstream platforms.
- 4.4.5 Due to the different cultural norms that operate in some niche spaces (whereby 'shitposting' and trolling are accepted, and even encouraged, forms of interaction) there is a lack of research into person-directed abuse, such as harassment and bullying, in niche spaces. As such, we cannot estimate prevalence of these behaviours.

5. Survey evidence on experiences of online abuse

5.1 Key findings from survey evidence

- 5.1.1 We estimate that 30-40% of people in the UK are exposed to online abuse.
- 5.1.2 We estimate that 10-20% of people in the UK have been targeted by abusive content online.
- 5.1.3 Experiences of online abuse vary across social media platforms. Facebook is consistently reported as the platform where individuals have experienced the most abuse. Our analysis of OxIS 2019 shows that experiences of abuse also vary by demographics:
 - 1. Ethnicity: Black people and those of 'Other' ethnicities are far more likely to be targeted by, and exposed to, online abuse than White and Asian people.
 - 2. Age: Younger people are more likely to be targeted by, and exposed to, online abuse.
 - 3. Gender: Surprisingly, our analysis of OxIS did not identify a substantial difference according to gender. However, we advise caution as other survey data suggests that gender plays an important role in shaping people's experiences of online abuse.

"A representative survey dedicated to understanding the experience of people in the UK of online abuse should be administered each year."

5.2 Oxford Internet Survey 2019 – new insights into people's experience of online abuse

5.2.1 We present original analysis of previously unpublished data from the Oxford Internet Institute's 2019 Oxford Internet Survey. We are grateful to the authors for giving us permission to use this data, and to Dr Grant Blank for making it available.¹⁰ OxlS is the longest-running academic survey of internet use in Great Britain and uses a multi-stage national probability sample of 2,000 people.

Current internet users, non-users and ex-users are included, which allows the data to be used for representative estimates of internet use, with a low margin of error. Two questions from the 2019 survey are pertinent to this report:

1. Have you seen cruel or hateful comments or images posted online?
2. Have you received obscene or abusive emails?

5.2.2 In total, 27% of respondents had seen cruel or hateful comments or images posted online. 10% of respondents had received obscene/abusive emails.

5.2.3 Headline figures can mask important differences in experiences of online abuse, based on respondents' demographics.¹¹ Ethnicity impacted respondents' experiences of online abuse. 7.7% of White respondents had received obscene/abusive emails, compared with 41.1% of Black respondents. A smaller difference also exists for seeing cruel/hateful content online. 26.6% of White respondents had viewed such content whilst 38.6% of Black respondents had. Asian and 'Other' respondents fell in between these two responses for both questions. Differences in experiences of online abuse according to ethnicity are shown in Figure 4.

5.2.4 Age impacted respondents' experiences of online abuse. Younger people are more likely to experience abuse. 41.2% of 18-30 year olds had seen cruel/hateful content online, compared with 7.4% of 76+ year olds. Age also impacted whether respondents had received obscene/abusive emails but the relationship was far weaker, ranging only from 13.1% of 18-30 year olds to 6.77% of 76+ year olds. Differences in experiences of online abuse according to age are shown in Figure 5.

¹⁰ Please see OxlS's website for more information about the survey, <https://oxis.oii.ox.ac.uk>.

¹¹ All of the differences identified here are statistically significant. Due to the sample size, we advise caution when interpreting differences which are less than 10 percentage points.

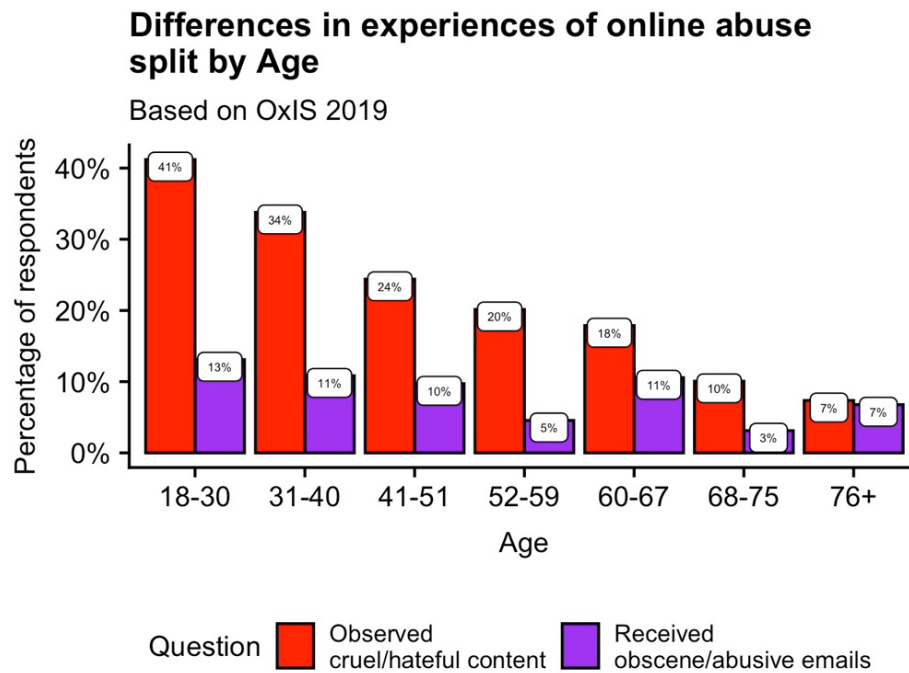


Figure 4: Experiences of online abuse, split by ethnicity (OxIS 2019 data)

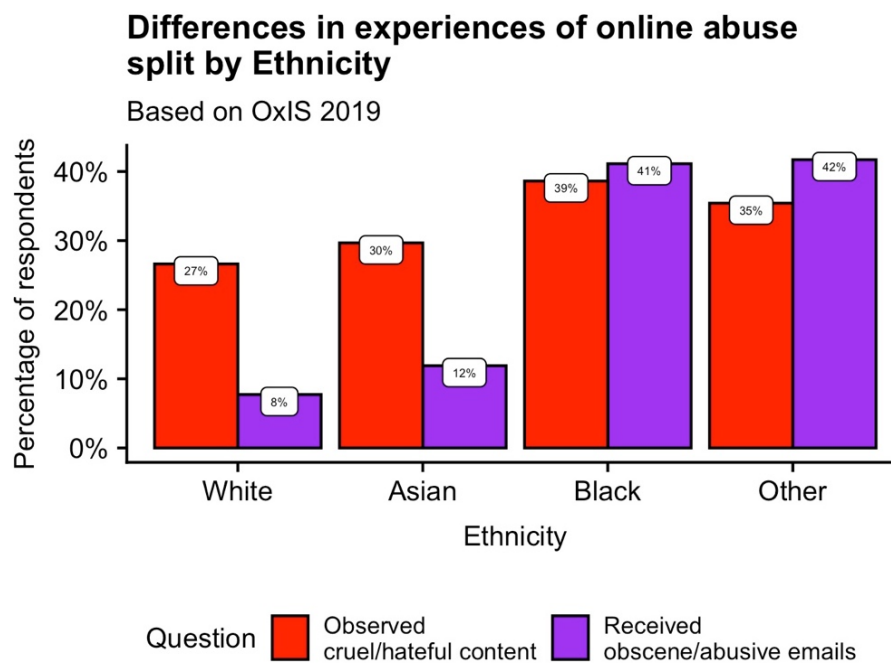


Figure 5: Experiences of online abuse, split by age (OxIS 2019 data)

5.2.5 Internet use impacted respondents' experiences of online abuse. 38.9% of users who 'constantly' go online had seen cruel/hateful content online, compared with just 5.2% of users who go online less often than once per week. More time spent online was also associated with being less likely to receive obscene/abusive emails, from 16.6% of users who 'constantly' go online to 2.2% of users who go online once per week. However, one exception is users who go online less often than once per week, of whom 21.5% had received obscene/abusive emails. Differences in experiences of online abuse according to internet use are shown in Figure 6.

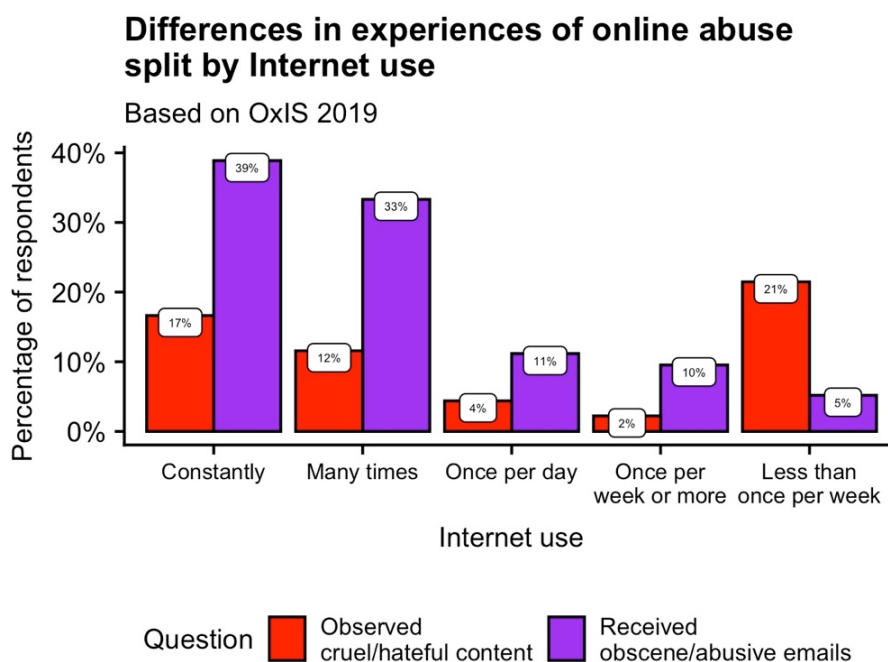


Figure 6: Experiences of online abuse, split by internet use (OxIS 2019 data)

5.2.6 40.7% of respondents who have a disability have seen cruel/hateful content online compared with 26.1% of respondents who do not have a disability. With regards to receiving obscene/abusive emails, the relationship is inverted, although the difference is far smaller and should be treated with caution: 8.5% of respondents with a disability have received such an email compared with 10.9% of those who do not. Differences in experiences of online abuse according to disability status are shown in Figure 7.

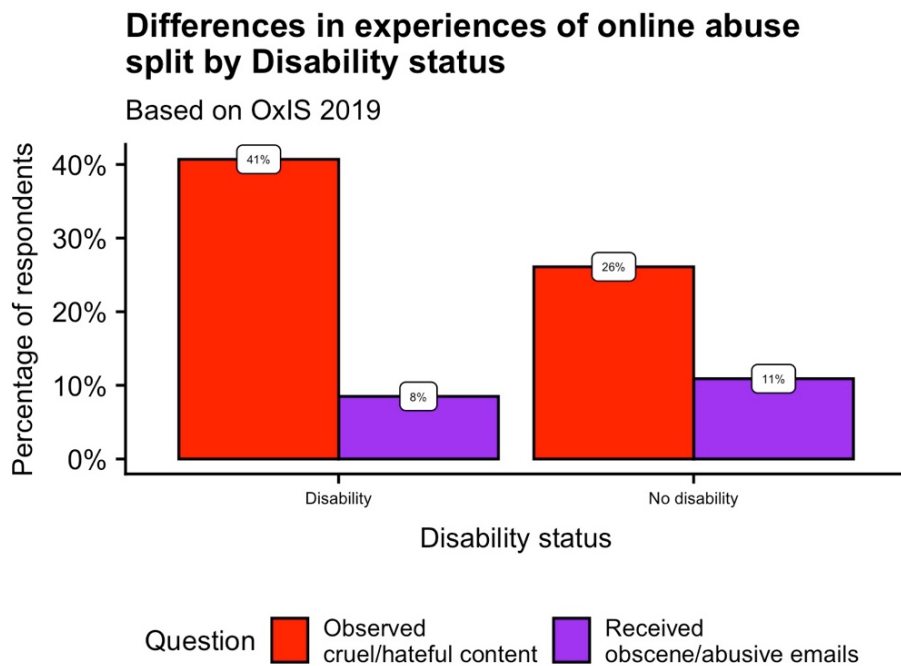


Figure 7: Experiences of online abuse, split by disability status (OxIS 2019 data)

- 5.2.7 Variables for other important aspects of identity, including income, education and gender, are available in the dataset. Surprisingly, our analysis did not show that people experienced different levels of online abuse according to their gender.
- 5.2.8 We encourage other researchers to request access to OxIS 2019 to conduct further analyses.

5.3 Survey review

- 5.3.1 Our extensive literature review identifies several surveys which investigate people's experiences of online abuse. The design, sampling, methodology, questions, timing and administration of surveys vary considerably and as such reported figures are often substantially different. Noticeably, few surveys focus on UK adult populations and we identify that an annual survey of UK adults' experiences of online abuse needs to be created, similar to existing surveys in New Zealand and America.
- 5.3.2 In Table 5 we show survey results for reported exposure to online abuse, based on 13 surveys. We estimate that 30-40% of all UK adults have been exposed to online abuse. This is also shown in Figure 8. Each of the surveys ask questions to very different groups, from representative samples of all UK adults to those which are far more specific, such as only including Lesbian, gay, bi and trans people. The different groups which are sampled are shown in the key of Figure 8.

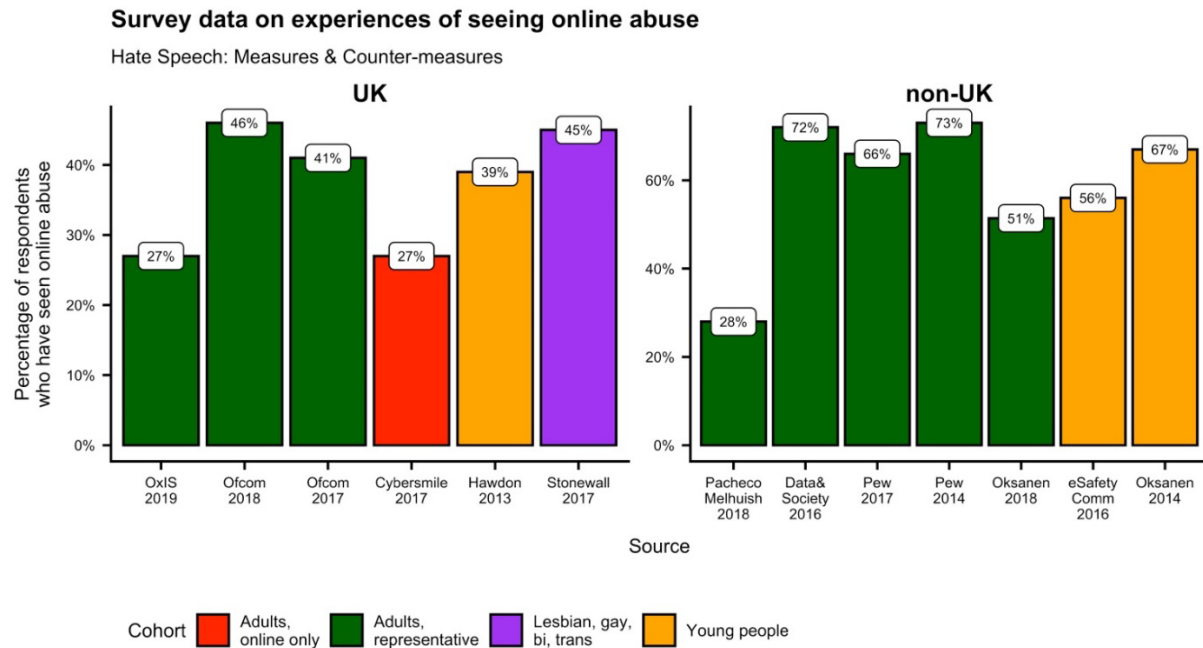


Figure 8: Survey data of experiences of seeing online abuse

- 5.3.3 In Table 6 we show survey results for reported experiences of being targeted by online abuse, based on 15 surveys. These surveys report very different figures, from 10% in OxIS (2019) to 85% in Galop (2017), which partly reflects how different groups systematically experience very different levels of personal abuse and partly the different survey designs.
- 5.3.4 We estimate that 10-20% of UK adults have been targeted by online abuse. This is shown in Figure 9. We anchor our estimate based on the findings of OxIS 2019 (10%) because this is the most representative survey of Internet use conducted in the UK, including both Internet users and non-Internet users, and is the most recently published result.
- 5.3.5 In Table 7 we report survey results which show the platforms where users report seeing online abuse, based on 3 surveys. The most widespread platform is Facebook, followed by Twitter and then either Instagram, Snapchat or WhatsApp. Note that these figures are not scaled by the number of users on each platform: there are far more users on Facebook than Snapchat and this may impact the percentage which have seen hate on there. This is shown in Figure 10.

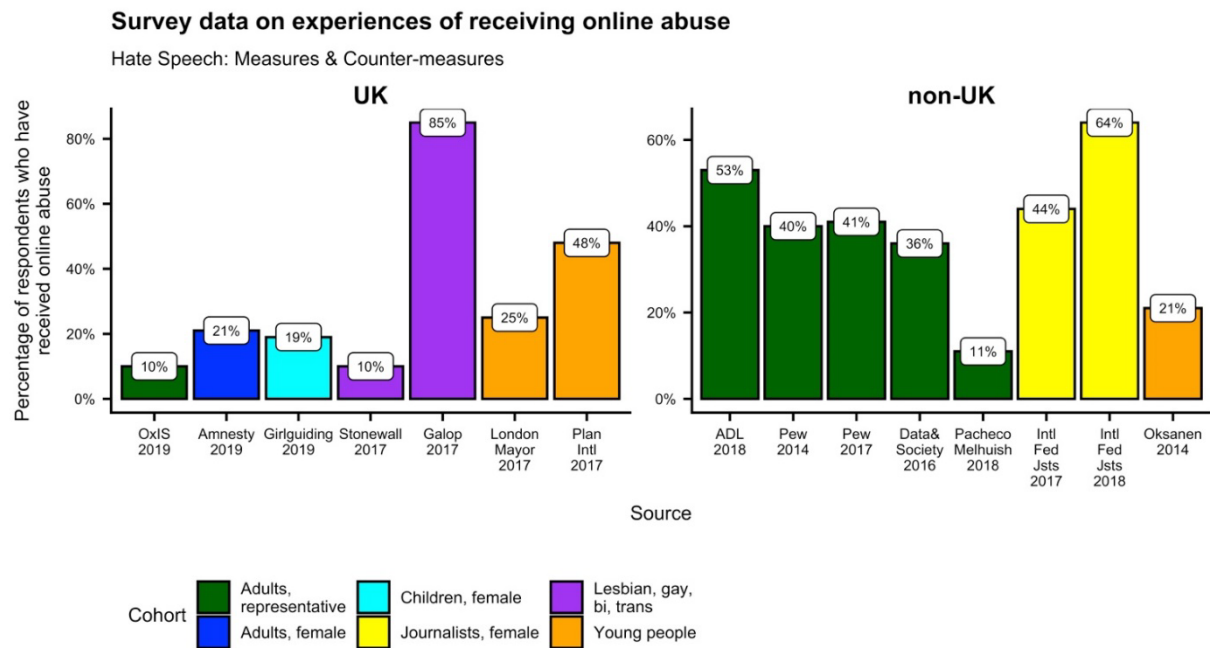


Figure 9: Survey data of experiences of receiving online abuse

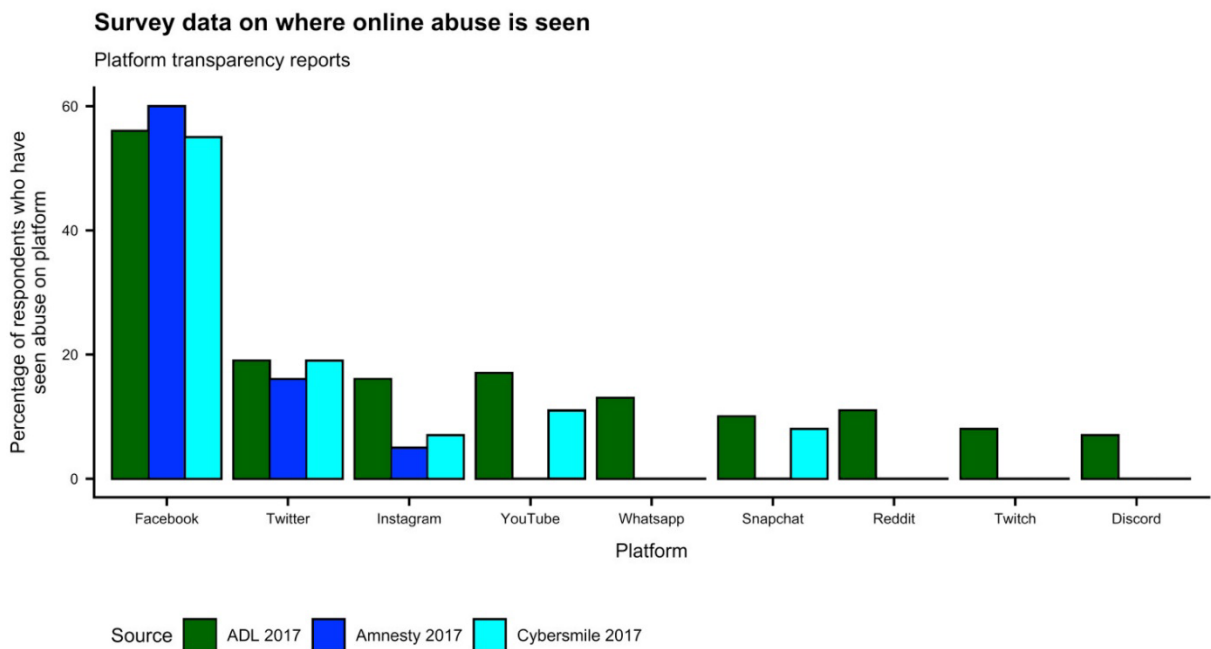


Figure 10: Survey data showing the platforms on which respondents exposed to abuse have seen it take place

Source	Details	Reported exposure to online abuse
OxIS (Blank and Dutton 2019)	Representative sample of 2,000 UK adults, including current, ex- and non- internet users. 2019.	27% had viewed cruel or hateful content.
Ofcom (Ofcom 2019a)	Representative sample of 1,882 UK adults. Conducted in 2018.	53% of internet users had viewed hateful content. 39% had 'sometimes' seen it and 14% had 'often' seen it. Given the percentage of adults who are internet users, 46% of all adults had viewed hateful content.
Ofcom (Ofcom 2018)	Representative sample of 1,875 UK adults. Conducted in 2017.	47% of internet users had viewed hateful content. 41% of all adults had viewed hateful content.
Cybersmile Foundation (Cybersmile 2017)	Sample of 4,231 UK adults, included a mix of regions, ages and genders. 2017.	27% had viewed bullying, abusive or harassing content.
Oksanen et al. (Oksanen et al. 2014)	Sample of 999 UK people aged 15 to 30 years-old, nationally representative for gender and geography within that age range. 2013.	39% had seen hateful or degrading writings or speech, which inappropriately attacked certain groups or individuals.
Stonewall (Stonewall 2017)	Sample of 5,375 lesbian, gay, bi and trans people from England, Scotland and Wales. 2017.	45% had witnessed homophobic, biphobic, and transphobic abuse or behaviour online that was directed towards other people.
Pacheco and Melhuish (Pacheco and Melhuish 2018)	Representative sample of 1,001 adults in New Zealand, weighted to ensure representation of minorities. 2018.	28% had seen or encountered online hate speech directed at someone else.
Data&Society (Data&Society 2016)	Nationally representative sample of 3,002 internet users aged 15 years or older in the USA, conducted in 2016.	72% had witnessed others experience abuse online.

Source	Details	Reported exposure to online abuse
Pew Research (Pew Research 2017)	Nationally representative sample of 4,248 American adults, conducted in 2017.	66% had witnessed harassing behaviour online directed at others.
Pew Research (Pew Research 2014)	Nationally representative sample of 2,849 American adult internet users, conducted in 2014.	73% had witnessed harassing behaviour online directed at others.
Oksanen et al. (Oksanen et al. 2018)	<p>Samples from 5 countries, stratified to represent the population of each country in terms of age, gender and geographic region. 2015, following the Paris terrorist attack.</p> <p>France, n = 2,113</p> <p>Spain, n = 1,661</p> <p>Finland, n = 1,003</p> <p>Norway, n = 1,013</p> <p>USA, n = 1,420</p>	<p>Respondents had viewed content that inappropriately attacked certain groups of people or individuals:</p> <p>France (36%)</p> <p>Spain (43%)</p> <p>Finland (57%)</p> <p>Norway (68%)</p> <p>USA (53%)</p> <p>Average: 51.4%</p>
eSafety Commissioner in Australia (Office of the eSafety Commissioner 2016)	Sample of 2,448 young people in Australia aged 12-17 from all states, conducted in 2016.	<p>56% had seen racist comments.</p> <p>53% had seen hateful comments directed against cultural or religious groups.</p>
Oksanen et al. (Oksanen et al. 2014).	Sample of 723 Finnish Facebook users, aged 15 to 18 years-old, including 472 females and 252 males. Other demographics broadly matched young people in Finland. 2013.	67% had been exposed to hate material online.

Table 5: Reported exposure to online abuse (results from 13 surveys)

Source	Details	Reported experiences of being targeted by online abuse
OxIS (Blank and Dutton 2019)	Representative sample of 2,000 UK adults, including current, ex- and non- internet users. 2019.	10% had received obscene/abusive emails.
Amnesty International (Amnesty International UK 2017a)	Sample of 500 UK women aged 18-55 years-old. 2017.	21% had experienced online abuse. They received: Threats of physical or sexual violence, (27%), Sexist or misogynistic comments, (47%) and Generally abusive language or comments, (69%).
Girlguiding (Girlguiding 2019)	Sample of 2,118 girls and young women aged 7-21 years-old. 2019.	19% of respondents aged 7-10 and 33% of respondents aged 11-21 had experienced mean comments from people online (See. p. 20).
Stonewall (Stonewall 2017)	Sample of 5,375 lesbian, gay, bi and trans people from England, Scotland and Wales. 2017.	10% of respondents had experienced homophobic, biphobic, or transphobic abuse online directed towards them personally in the last month. Including: 26% of people who identify as transsexual; 20% of black, Asian and minority ethnic LGBT people; 26% of non-binary LGBT people (See p. 19).
Galop (Galop 2017)	Sample of 271 LGBT+ people (Galop 2017). Participants were recruited through a network of LGBT+ activists, individuals and professionals and the sample was not representative.	85% of respondents had experienced online abuse more than once. 59% of respondents had experienced online abuse six or more times.
The Mayor's Office of Policing and Crime (MOPAC 2018)	Sample of 7,832 young people in London aged 11-16 years old. Respondents included a mix of genders, ethnicities and geographic boroughs. 2018.	25% of students in years 10 and 11 (aged approximately 14 to 16) had experienced someone saying mean things or bullying online.

Source	Details	Reported experiences of being targeted by online abuse
Plan International UK (Opinium)	Sample of 1,002 young people in the UK, aged 11-18. 2017.	48% of female respondents and 40% of males had experienced some form of harassment or abuse on social media.
The Anti-Defamation League (ADL 2019)	Nationally representative sample of 1,134 American adults. Data was weighted on the basis of age, gender identity, race, census region and education. 2018.	53% of respondents had experienced some type of online harassment. This includes: offensive name calling (41%), someone trying to purposefully embarrass them (33%), and severe online harassment, which includes sexual harassment, stalking, physical threats, and sustained harassment (37%).
Pacheco and Melhuish (Pacheco and Melhuish 2018)	Nationally representative sample of 1,001 New Zealand adults (Pacheco and Melhuish 2018, 3). 2017.	11% of respondents had been personally targeted with online hate speech in the prior year. 4% had received hate once, 5% between 2 and 4 times, and 1% 5 or more times.
Data&Society (Data&Society 2016)	Nationally representative sample of 3,002 internet users aged 15 years or older living in the USA. (Data&Society 2016).	36% of respondents had personally experienced a direct form of harassment, such as being called offensive names, being physically threatened and being stalked.
Pew Research (Pew Research 2017)	Nationally representative sample of 4,248 American adults. 2017.	41% of respondents had experienced harassing behaviour online. This includes: offensive name-calling (27%), purposeful embarrassment (22%) severe forms of harassment online, including physical threats, harassments over a sustained period, sexual harassment and stalking (10%).

Source	Details	Reported experiences of being targeted by online abuse
Pew Research (Pew Research 2014)	Nationally representative sample of 2,849 American adult internet users. 2014.	40% of respondents had experienced harassing behaviour online. 22% had experienced less severe forms (being called offensive names and being purposefully embarrassed) and 18% had experienced more severe forms (being physical threatened, stalked, harassed for a sustained period of sexually harassed).
The International Federation of Journalists (IFJ 2017)	Sample of ~400 women journalists from 50 countries. 2017.	44% had experienced online abuse.
The International Federation of Journalists (IFJ 2018)	Sample of 267 journalists, of which 162 were female. The responses of males have not been published. 2018.	64% of female respondents reported they had experienced online abuse. This includes: sexist insults (48%), threats to them or their family (20%) and threats of violence (19%).
Oksanen et al. (Oksanen et al. 2014).	Sample of 723 Finnish Facebook users, aged 15 to 18 years-old, including 472 females and 252 males. Other demographics broadly matched young people in Finland. 2013 (Oksanen et al. 2014).	21% of respondents had been targeted by hate material online.

Table 6: Reported experiences of being targeted by online abuse (15 surveys)

Survey	Facebook	Twitter	Instagram	Snapchat	Whatsapp
Amnesty International (Amnesty International UK 2017a)	60%	16%	5%	N/A	N/A
ADL (ADL 2019)	56%	19%	16%	10%	13%
Cybersmile (Cybersmile 2017)	55%	19%	7%	8%	N/A

Survey	Reddit	YouTube	Twitch	Discord
Amnesty International (Amnesty International UK 2017a)	N/A	N/A	N/A	N/A
ADL (ADL 2019)	11%	17%	8%	7%
Cybersmile (Cybersmile 2017)	N/A	11%	N/A	N/A

Table 7: Reported platforms on which respondents exposed to abuse have seen it take place

Term	Acronym
BME	Black and minority ethnic
ADL	Anti-Defamation League
CST	Community Security Trust
Intl	International
Fed	Federation

Term	Acronym
Jsts	Journalists
DCMS	Department of Digital, Culture, Media and Sport
Ofcom	Office of Communications
OxIS	Oxford Internet Survey

Table 8: Table of Acronyms and abbreviations

Acknowledgements

We would like to thank Beth Wood, Mark Burey, Harry Thompson and Amit Mulji for their help in finalising the report.

References

Research publications and reports

- Abrams, Dominic, Hannah Swift, and Diane Houston. 2018. *Developing a National Barometer of Prejudice and Discrimination in Britain*. London.
<https://www.equalityhumanrights.com/sites/default/files/national-barometer-of-prejudice-and-discrimination-in-britain.pdf>.
- ADL. 2019. *Online Hate and Harassment: The American Experience*. New York.
<https://www.adl.org/onlineharassment>.
- Amnesty International UK. 2017a. Social Media Can Be a Dangerous Place for Women. London.
- . 2017b. *Tackling Hate Crime in the UK*. London.
<https://www.amnesty.org.uk/files/Against-Hate-Briefing-AIUK.pdf>.
- . 2017c. Amnesty Global Insights Unsocial Media: Tracking Twitter Abuse against Women MPs. London.
- Bakalis, Chara. 2018. "Rethinking Cyberhate Laws." *Information and Communications Technology Law* 27(1): 86–110.
- Benesch, Susan. 2012. *Dangerous Speech: A Proposal to Prevent Group Violence*. New York.
- Blank, Grant. 2017. "The Digital Divide Among Twitter Users and Its Implications for Social Research." *Social Science Computer Review* 35(6): 679–97.
- Blank, Grant, and William H. Dutton. 2019. *OxIS, Perceived Threats to Privacy Online: The Internet in Britain*. Oxford. <https://oxis.oii.ox.ac.uk/wp-content/uploads/sites/43/2019/09/OxIS-report-2019-final-digital-PDFA.pdf>.
- Bliuc, Ana Maria, Nicholas Faulkner, Andrew Jakubowicz, and Craig McGarty. 2018. "Online Networks of Racial Hate: A Systematic Review of 10 Years of Research on Cyber-Racism." *Computers in Human Behavior* 87(1): 75–86.
<https://doi.org/10.1016/j.chb.2018.05.026>.

- Bowman-Grieve, Lorraine. 2009. "Exploring Stormfront: A Virtual Community of the Radical Right." *Studies in Conflict and Terrorism* 32(11): 989–1007.
<http://www.tandfonline.com/toc/uter20/32/11>.
- Brown, Alexander. 2017. "What Is Hate Speech? Part 1: The Myth of Hate." *Law and Philosophy* 36(4): 419–68.
- . 2018. "What Is so Special about Online (as Compared to Offline) Hate Speech?" *Ethnicities* 18(3): 297–326.
- Butler, Judith. 1997. *Excitable Speech: A Politics of the Performative*. New York: Routledge.
- Chandrasekharan, Eshwar et al. 2017. "You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech." *Proceedings of the ACM on Human-Computer Interaction* 1(2): 1–22.
- Cihon, Peter, and Taha Yasseri. 2016. "A Biased Review of Biases in Twitter Studies on Political Collective Action." *Frontiers in Physics* 4(1): 1–10.
<http://arxiv.org/abs/1605.04774>.
- Cohen-Almagor, Raphael. 2011. "Fighting Hate and Bigotry on the Internet." *Policy & Internet* 3(3): 89–114. <http://doi.wiley.com/10.2202/1944-2866.1059>.
- Commission for Countering Extremism. 2019. *Challenging Hateful Extremism*. London.
- CST. 2019. *Antisemitic Incidents Incidents*. London. www.cst.org.uk.
- Cybersmile. 2017. *The Cybersmile Foundation Stop Cyberbullying Day Survey 2017*. London.
- Dadvar, Maral, Dolf Trieschnigg, Roeland Ordelman, and Franciska De Jong. 2013. "Improving Cyberbullying Detection with User Context." In *Lecture Notes in Computer Science*, , 693–96.
- Data&Society. 2016. *Online Harrassment, Digital Abuse, and Cyberstalking in America*. https://www.datasociety.net/pubs/oh/Online_Harassment_2016.pdf.
- Davidson, Julia et al. 2019. *Adult Online Hate, Harassment and Abuse: A Rapid Evidence Assessment*. London: UK Council for Internet Safety.
- Demos. 2017. *Anti-Islamic Hate on Twitter*. London.
- Eatwell, Roger. 2006. "Community Cohesion and Cumulative Extremism in Britain." *Political Quarterly* 77(2): 204–16.
- Echikson, William, and Olivia Knodt. 2018. *CEPS Policy Insight Germany's NetzDG: A Key Test for Combatting Online Hate*. Brussels.
<https://www.ceps.eu/publications/germany's-netzdg-key-test-combatting-online-hate>.

- ECRI. 2015. ECRI General Policy Recommendation No. 9 on Combating Hate Speech. Strasbourg.
- Figea, Leo, Lisa Kaati, and Ryan Scrivens. 2016. "Measuring Online Affects in a White Supremacy Forum." *IEEE International Conference on Intelligence and Security Informatics: Cybersecurity and Big Data, ISI 2016*: 85–90.
- Founta, Antigoni-maria et al. 2018. "Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior." In *ICWSM*, , 1–11.
- FRA. 2018. *Hate Crime Recording and Data Collection Practice across the EU*. Luxembourg: European Agency for Fundamental Rights.
- Galop. 2017. Galop Online Hate Crime Report. London.
- Girlguiding. 2019. *Girls' Attitudes Survey 2019*. London.
- Golbeck, Jennifer et al. 2017. "A Large Labeled Corpus for Online Harassment Research." In *Proceedings of the ACM Conference on Web Science*, , 229–33.
- Gorrell, Genevieve et al. 2018. Arxiv preprint arXiv1804.01498 *Online Abuse of UK MPs in 2015 and 2017: Perpetrators, Targets, and Topics*. <http://arxiv.org/abs/1804.01498>.
- Greenwood, Mark A et al. 2019. Arxiv preprint arXiv1904.11230 *Online Abuse of UK MPs from 2015 to 2019*.
- Hine, Gabriel Emile et al. 2017. "Keks, Cucks and God Emperor Trump: A Measurement Study of 4Chan's Politically Incorrect Forum and Its Effect on the Web." In *ICWSM*, , 92–101. <http://arxiv.org/abs/1610.03452> (March 11, 2019).
- HM Government. 2017. Home Affairs Committee Report on Hate Crime: Abuse, Hate and Extremism Online. London.
- . 2019. *Online Harms White Paper*. London: Department of Digital, Culture, Media and Society.
- Home Office. 2016. Action Against Hate: The UK Government's Plan for Tackling Hate Crime. London.
- . 2017. Hate Crime, England and Wales, 2016/17 Statistical Bulletin. London.
- . 2018. *Hate Crime, England and Wales, 2017/18 Statistical Bulletin*. London. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/467366/hosb0515.pdf.
- IFJ. 2017. *IFJ Global Survey 2017*. Brussels, Belgium.
- . 2018. *IFJ Global Survey 2018*. Brussels, Belgium.

- Keck, Thomas M. 2016. "Hate Speech and Double Standards." *Constitutional Studies* 1(1): 95–121. <https://ssrn.com/abstract=2703484>.
- Krasodonski-Jones, Alex. 2017. *Signal and Noise: Can Technology Provide a Window into the New World of Digital Politics in the UK?* London.
- Mathew, Binny, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2018. "Spread of Hate Speech in Online Social Media." In *Proceedings of the 10th ACM Conference on Web Science*, Boston. <http://arxiv.org/abs/1812.01693>.
- Matsuda, Mary, Charles Lawrence, Richard Delgado, and Kimberlé Crenshaw. 1993. *Words That Wound: Critical Race Theory, Assaultive Speech and the First Amendment*. New York: Routledge.
- Miller, Carl, Alex Krasodonski-jones, and Jack Dale. 2016. *From Brussels to Brexit : Islamophobia, Xenophobia , Racism and Reports of Hateful Incidents on Twitter Research*. London.
- Miller, Carl, Josh Smith, Jack Dale, and Social Media. 2016. *Islamophobia on Twitter: March to July 2016*. London.
- Modood, T. et al. 2006. "The Danish Cartoon Affair: Free Speech, Racism, Islamism, and Integration." *International Migration* 44(5): 3–62.
- MOPAC. 2018. *MOPAC Evidence and Insight Youth Voice Survey Report 2018*. London.
- Nadal, Kevin L et al. 2012. "Subtle and Overt Forms of Islamophobia: Microaggressions toward Muslim Americans." *Journal of Muslim Mental Health* 6(2): 15–37.
- Nash, Victoria. 2019. "Revise and Resubmit ? Reviewing the 2019 Online Harms White Paper." *Journal of Media Law* 0(0): 1–10. <https://doi.org/10.1080/17577632.2019.1666475>.
- Northern Ireland Statistics and Research Agency. 2019. *Incidents and Crimes with a Hate Motivation Recorded by the Police in Northern Ireland*. Belfast.
- Ofcom. 2018. *Adults' Media Use and Attitudes Report*. London.
- . 2019a. *Adults' Media Use and Attitudes Report*. London.
- . 2019b. *Internet Users' Concerns about and Experience of Potential Online Harms*. London. <https://www.ofcom.org.uk/research-and-data/internet-and-on-demand-research/internet-use-and-attitudes/internet-users-experience-of-harm-online-2019>.
- . 2019c. *Use of AI in Online Content Moderation 2019*. London.

- Office of the eSafety Commissioner. 2016. *Young People's Experience with Online Hate, Bullying and Violence*. Canberra. <https://www.esafety.gov.au/education-resources/iparent/online-hate-infographic>.
- Oksanen, Atte et al. 2014. "Exposure to Online Hate among Young Social Media Users." *Sociological Studies of Children and Youth* 18(1): 253–73.
- . 2018. "Perceived Societal Fear and Cyberhate after the November 2015 Paris Terrorist Attacks." *Terrorism and Political Violence* 00(00): 1–20. <https://doi.org/10.1080/09546553.2018.1442329>.
- OSCE ODIHR. 2018. "ODIHR's Annual Reporting on Hate Crime." *OSCE*.
- Pacheco, Edgar, and Neil Melhuish. 2018. *NetSafe Online Hate Speech: A Survey on Personal Experiences and Exposure Among Adult New Zealanders*. Wellington.
- Pettigrew, Thomas F, and R. W. Meertens. 1995. "Subtle and Blatant Prejudice in Western Europe." *European Journal of Social Psychology* 25(1): 57–75.
- Pew Research. 2014. *Online Harassment*. New York.
- . 2017. *Online Harassment*. New York.
- . 2019. "Share of U.S. Adults Using Social Media, Including Facebook, Is Mostly Unchanged since 2018." *Pew Research Center: Fact Tank*. <https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/> (June 9, 2019).
- Robertson, Adi. 2013. "Facebook Users Have Uploaded a Quarter Trillion Photos since the Site's Launch." <http://www.theverge.com/2013/9/17/4741332/facebook-users-have-uploaded-a-quarter-trillion-photos-since-launch>.
- Salminen, Joni et al. 2018. "Online Hate Interpretation Varies by Country, But More by Individual." In *Proceedings of SNAMS*, , 1–7.
- Schmidt, Anna, and Michael Wiegand. 2017. "A Survey on Hate Speech Detection Using Natural Language Processing." In *International Workshop on NLP for Social Media*, Valencia, Spain, 1–10.
- Scotland Prosecution Service. 2017. *Government of Scotland Hate Crime in Scotland*. Edinburgh. <http://dx.doi.org/10.1016/bs.ampbs.2017.04.001><http://dx.doi.org/10.1016/j.arabjc.2013.08.010><http://dx.doi.org/10.1016/j.chemosphere.2013.01.075><http://www.pnas.org/cgi/doi/10.1073/pnas.0308555101><http://www.treemediation.com/technical/phytoremed>.
- Scottish Government. 2018. *Scottish Crime and Justice Survey 2017/18*. Edinburgh.

- SELMA. 2019. Hacking Online Hate : Building an Evidence Base for Educators.
- Simpson, Robert Mark. 2018. "Regulating Offense, Nurturing Offense." *Politics, Philosophy and Economics* 17(3): 235–56.
- Stephens-Davidowitz, Seth. 2019. Hidden Hate: What Google Searches Tell Us about Antisemitism Today. London.
- Stonewall. 2017. LGBT in Britain: Hate Crime and Discrimination. London.
- Tell Mama. 2018. *Beyond the Incident*. London.
- The Law Commission. 2018. Abusive and Offensive Online Communications: A Scoping Report. London: The Law Commission.
- Tworek, Heidi, and Paddy Leerssen. 2019. *An Analysis of Germany's NetzDG Law*. Pennsylvania.
- Vidgen, Bertie et al. 2019. "Challenges and Frontiers in Abusive Content Detection." In *3rd Workshop on Abusive Language Online*,.
- Waseem, Zeerak, Thomas Davidson, Dana Warmusley, and Ingmar Weber. 2017. "Understanding Abuse: A Typology of Abusive Language Detection Subtasks." In *1st Workshop on Abusive Language Online*, , 78–84. <http://arxiv.org/abs/1705.09899>.
- Wiegand, Michael, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. "Detection of Abusive Language: The Problem of Biased Datasets." In *NAACL-HLT*, , 602–8.
- Williams, Matthew, and Pete Burnap. 2016. "Cyberhate on Social Media in the Aftermath of Woolwich: A Case Study in Computational Criminology and Big Data." *British Journal of Criminology* 56(1): 211–38.
- Williams, Matthew, and Olivia Pearson. 2016. *Hate Crime and Bullying in the Age of Social Media*. Cardiff.
- Wulczyn, Ellery, Nithum Thain, and Lucas Dixon. 2017. "Ex Machina: Personal Attacks Seen at Scale." In *Proceedings of the International World Wide Web Conference*, , 1391–99. <http://arxiv.org/abs/1610.08914>.
- Zannettou, Savvas et al. 2018. "What Is Gab? A Bastion of Free Speech or an Alt-Right Echo Chamber? ACM Reference Format." In *Proceedings of the International World Wide Web Conference*, ACM, 1–8. <https://doi.org/10.1145/3184558.3191531> (March 11, 2019).

News articles, blogs and websites

BBC News. 2018. "Focus on Violent Crime Not Misogyny, Says Police Chief." *BBC News*.

CBS News. 2019. "Facebook Wants You to Post More Personal Updates, Not Just News." *CBS News*.

DMR. 2019. "80 Amazing Reddit Statistics and Facts (2019) | By the Numbers." <https://expandedramblings.com/index.php/reddit-stats/> (November 20, 2019).

Facebook. 2019. "Facebook Community Standards Enforcement Report." *Facebook Transparency* (April).

Hertfordshire Constabulary. 2018. Hertfordshire Constabulary: Annual Management Statement. Hertfordshire.

Human Rights Watch. 2018. "Germany: Flawed Social Media Law." *Human Rights Watch*. <https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law>.

Merseyside Police. 2019. "Welcome to the Merseyside Police Website."

Metropolitan Police. 2017. "Metropolitan Police Service Business Plan 2017-18." : 1–72. <https://www.facebook.com/metpoliceuk/>.

Reddit. 2019. "Reddit By The Numbers." *Reddit*. <https://www.redditinc.com> (November 20, 2019).

The Guardian. 2016. "The Dark Side of Guardian Comments." *The Guardian*.

———. 2019a. "8chan: Owner of Extremist Site Lashes out as Scrutiny Intensifies." *The Guardian*.

———. 2019b. "France Online Hate Speech Law to Force Social Media Sites to Act Quickly." *The Guardian*: 24–25. https://www.theguardian.com/world/2019/jul/09/france-online-hate-speech-law-social-media?CMP=Share_iOSApp_Other.

The Metropolitan Police. 2016. Metropolitan Police: Freedom of Information Request, 2006/07-2015/16. London.

The Washington Post. 2018. "Transcript of Mark Zuckerberg's Senate Hearing." *The Washington Post*.

Tubics. 2018. *How Many YouTube Channels Are There?* San Francisco. <https://www.tubics.com/blog/number-of-youtube-channels/>.

Twitter. 2019a. *Q1 2019 Letter to Shareholders*. San Francisco. https://s22.q4cdn.com/826641620/files/doc_financials/2019/q1/Q1-2019-Shareholder-Letter.pdf.

- . 2019b. “Twitter Rules Enforcement.” *Twitter Transparency report*.
<https://transparency.twitter.com/en/twitter-rules-enforcement.html> (November 20, 2019).
- West Midlands Police. 2019. “West Midlands Police.”
- YouTube. 2019a. *The Four Rs of Responsibility*. San Francisco.
- . 2019b. “YouTube Community Guidelines Enforcement.” *YouTube Community Standards*. http://www.youtube.com/t/community_guidelines (November 20, 2019).