How much online abuse is there?

A systematic review of evidence for the UK

Policy Briefing – Summary

Bertie Vidgen, Helen Margetts, Alex Harris

The Alan Turing Institute

Public Policy Programme
Hate Speech: Measures and Counter Measures

Contents

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.

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).

Assessing the prevalence of online abuse is a difficult task. Nonetheless, 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[1] and a survey of 500 women in the UK showed that 61% of respondents believed online abuse/harassment of women is common and 47% stated that current laws are inadequate. A recent report from the Commission for Countering Extremism found that 56% of the public believe 'a lot more' should be done to counter extremism online.[2]

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, 2019).

[1] https://www.ofcom.org.uk/__data/assets/pdf_file/0028/149068/online-harms-chart-pack.pdf
[2] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/836538/Challenging_Hateful_Extremism_report.pdf

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[3], 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.

[3] For more information, please see our website, https://www.turing.ac.uk/research/research-projects/hate-speech-measures-and-counter-measures

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)[4], 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.

[4] 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.

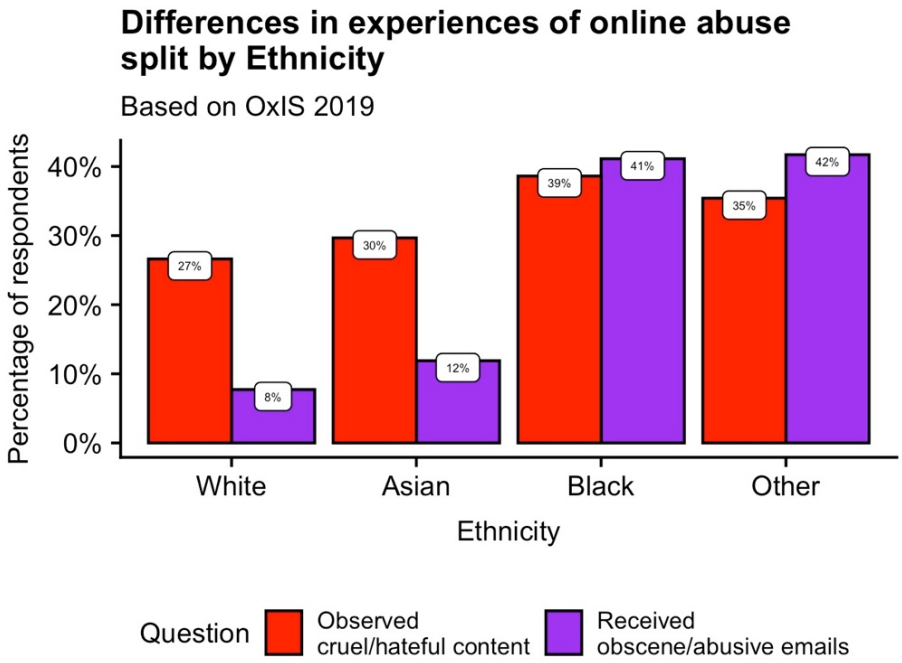

Figure 1: Experiences of online abuse, split by ethnicity (OxIS 2019 data)

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.[5]

[5] For one noticeable exception, see (Stephens-Davidowitz 2019).

Conclusions

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).

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).

As more time and resources are deployed to monitor, understand and counter hate[6] it is crucial that we develop a better understanding of the scale of the problem. This policy report, in both summary and full form, provides a first step in building a robust evidence base to understand online abuse. In our future work, The Alan Turing Institute's 'Hate Speech: Measures and Counter-measures' project will continue to build out more resources for all researchers, policymakers, civil society actors and industry practitioners working in this space.

[6] See: https://www.gov.uk/government/news/uk-to-help-develop-new-tech-to-stop-sharing-of-terrorist-content and https://www.london.gov.uk/press-releases/mayoral/mayor-launches-unit-to-tackle-online-hate-crime

References

Amnesty International UK. (2017). *Tackling Hate Crime in the UK*. London. Retrieved from https://www.amnesty.org.uk/files/Against-Hate-Briefing-AIUK.pdf

Brown, A. (2018). What is so special about online (as compared to offline) hate speech? *Ethnicities*, *18*(3), 297–326. https://doi.org/10.1177/1468796817709846

Davidson, J., Livingstone, S., Jenkins, S., Gekoski, A., Choak, C., & Phillips, K. (2019). *Adult Online Hate, Harassment and Abuse: a rapid evidence assessment*. London: UK Council for Internet Safety.

Eatwell, R. (2006). Community Cohesion and Cumulative Extremism in Britain. *Political Quarterly*, *77*(2), 204–216.

ECRI. (2015). *ECRI General Policy Recommendation No. 9 on combating hate speech*. Strasbourg.

HM Government. (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.

Home Office. (2018). *Hate Crime, England and Wales, 2017/18 Statistical Bulletin*. London. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/467366/hosb0515.pdf

Nash, V. (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

Ofcom. (2019). *Use of AI in online content moderation 2019*. London.

SELMA. (2019). *Hacking Online Hate : Building an Evidence Base for Educators*.

Stephens-Davidowitz, S. (2019). *Hidden hate: What Google searches tell us about antisemitism today*. London.

The Law Commission. (2018). *Abusive and Offensive Online Communications: A scoping report*. London: The Law Commission.

The Washington Post. (2018, April 11). Transcript of Mark Zuckerberg's Senate hearing. *The Washington Post*.

Vidgen, B., Tromble, R., Harris, A., Hale, S., Nguyen, D., & Margetts, H. (2019). Challenges and frontiers in abusive content detection. In *3rd Workshop on Abusive Language Online*.