The
Alan Turing
Institute

Understanding online hate

VSP Regulation and the broader context

Bertie Vidgen
Emily Burden
Helen Margetts

The Alan Turing Institute

Understanding Online Hate

VSP regulation and the broader context

Bertie Vidgen

Emily Burden

Helen Margetts

February 2021

CONTENT WARNING: This report contains examples of online hate and, as a result, contains a range of language which may cause offence.

Introduction

This report aims to contribute to our understanding of online hate in the context of the requirements of the revised Audiovisual Media Services Directive (AVMSD) [1] for Video Sharing Platforms (VSPs)[2] to protect the general public from incitement to hatred or violence. However, online hate is complex and it can only be fully understood by considering issues beyond the very specific focus of these regulations. Hence, we draw on recent social and computational research to consider a range of points outside VSP regulations, such as the impact, nature and dynamics of online hate. For similar reasons, we have considered expressions of hate across a range of online spaces, including VSPs as well as other online platforms. In particular, we have closely examined how online hate is currently addressed by industry, identifying key and emerging issues in content moderation practices.[3] Our analyses will be relevant to a range of experts and stakeholders working to address online hate, including researchers, platforms, regulators and civil society organisations.

To guide this report, we have used a definition of online hate from forthcoming work by The Alan Turing Institute. Our findings and recommendations are based on this

[1] UK Legislation, *The Audiovisual Media Services Regulations 2020* (London: UK, 2020). Available at: https://www.legislation.gov.uk/uksi/2020/1062/made

[2] A VSP is an online service provided to members of the public, whose principal purpose or essential functionality is to offer videos which are uploaded by its users, rather than the service provider who does not have general control over what videos are available but does have general control over the manner in which videos are organised, for example automatically or by way of algorithms. For further explanation of the legal criteria, see https://www.ofcom.org.uk/__data/assets/pdf_file/0021/205167/regulating-vsp-guide.pdf.

[3] The report does not purport to determine if specific services are within the scope of VSP regulations, nor does it intend to assess if specific services meet their compliance requirements.

definition, and not on the legal references in the AVMSD to protect users from incitement to hatred or violence. Our definition is: [4]

> "Online hate speech is a communication on the Internet which expresses prejudice against an identity. It can take the form of derogatory, demonising and dehumanising statements, threats, identity-based insults, pejorative terms and slurs. Online hate speech involves:
>
> 1. A medium for the content such as text, images, video, audio, and gifs;
> 2. A perpetrator (the person who creates or shares the hateful content);
> 3. An actual or potential audience (anyone who is or who could be exposed to or targeted by the content);
> 4. A communicative setting (e.g., private messaging apps, online forums, comment sections or broadcast-style social media platforms)."

Summary of key findings

Characterising, defining and regulating online hate poses a range of challenges. Across society there is a lack of consensus about how online hate should be defined, particularly regarding what 'edge case' content should be considered hateful. Key differences exist between online and offline forms of hate in terms of their nature, dynamics and prevalence. Nonetheless, policymakers and researchers alike have emphasised the severity of hateful communications online and it is widely accepted that they are **as capable of inflicting harm on victims as offline hate**.

A large and growing body of research, spanning social, behavioural and computational scientific fields, has analysed online hate**. Key findings from previous research include**:

1. Assessing the prevalence of online hate is difficult due to the lack of appropriate data and robust measurement tools.

[4] Based on forthcoming work from The Alan Turing Institute, developed by Bertie Vidgen, Josh Cowls and Helen Margetts.

2. Online hate is perpetrated by a wide range of actors, including both lone individuals (such as 'bedroom' trolls) and individuals affiliated with hate groups, such as white supremacist organisations.
3. The prevalence and dynamics of online hate vary across platforms. In general, there is far less overt hate on more mainstream platforms compared with niche alternative spaces.
4. Online hate spreads through networks and some hateful content can quickly reach very large audiences.
5. Online hate is an event-driven landscape, affected by elections, terror attacks and stories in the news.
6. There are close links between online and offline forms of hate, although identifying the direction of causality is often difficult.
7. Online hate exhibits cross-platform dynamics, such as how some users migrate between different platforms following influencer bans.
8. Experiences of being exposed to and targeted by online hate vary according to many factors, including age, gender, ethnicity and religion.

To better understand different types of online hate, we analyse both the **substance** of content and how it is **articulated.**

- **Substance captures what the hate expresses:** Hate can be expressed, inspired and incited through more than just directly threatening or violent language. We categorise online hate into four types of speech: **threatening, inciting, demonising and animosity**. The biggest challenge in tackling online hate is addressing the 'grey area' of animosity, which is often ambiguous and highly contested. The substance of hate depends on what it contains as well as the context in which it is produced. How it is received will depend in part on the subjective outlook of the audience.
- **Articulation captures how the hate is expressed:** hate can be articulated both **covertly** and **overtly**. Overt forms of hate are usually more aggressive and can include 'amplifying' elements, such as swear words. Covert forms are harder to

identify and may intentionally be expressed in a misleading way through ‘obfuscation’. Such content will be harder for content moderation systems to detect. Some forms of hate will be difficult even for trained experts to recognise due to the use of ‘code words’ and complex forms of language.

The combination of what online hate expresses (its substance) and how it is expressed (its articulation) produces its **hazard**, which is the potential of the content to inflict harm in a given context. The actual harm that hateful content inflicts is a product of its hazard and **influence**, which we define as its resonance and reach. Hateful content created by more authoritative speakers, seen by more people, and by audiences which are more susceptible to its message, is likely to have more influence and therefore more hazard.

We identify four types of language which capture how the hazard and influence of hateful content intersect: **Dangerous speech** is highly hazardous content which has substantial influence, impacting many people. **Bedroom trolling** is content that might be equally hazardous but has little influence. **Benign viral content** is non-hateful content which is seen by many people. **Everyday talk** is content which contains no hate and has little impact.

The harm in hate speech has been keenly debated by academics and civil society activists. To provide clarity we identify **seven non-exhaustive ways in which online hate can inflict harm:**

1. Immediate distress and emotional harm on victims;
2. Long-term mental health effects;
3. Long-term impact on victims’ behaviour;
4. Negative impact on individuals’ willingness to engage in civic and public discourse;
5. Motivating and enabling offline hate attacks;
6. Motivating and enabling online attacks, such as hacking or scamming members of a targeted group;
7. The negative implications for social justice and fairness**.**

Tackling the harm inflicted by online hate is difficult work. Most online platforms, including VSPs, address online hate by creating content moderation systems. These are socio-technical systems, comprising infrastructure, people, technology and policies. We identify **five desirable features of moderation systems**: (1) High-performing, (2) fair, (3) robust, (4) explainable and (5) scalable.

To help understand how an effective content moderation system could be designed, we outline four core activities which need to be carried out:

1) **Characterise online hate**: Provide a clear definition of online hate, construct a typology (if needed) and outline guidelines, including examples, rationales and principles. The 'tricky' issues in hate should be engaged with to establish where the line falls between hate and non-hate.
2) **Implement strategies for identifying online hate.** This will vary across platforms, depending on their expertise, infrastructure and budget. Broadly, three planks form the basis of most content moderation processes for identifying online hate: User reports, AI, and human review. **There are inherent limitations in using humans to moderate flagged content** (i.e. it is time consuming, expensive, can be inconsistent, and has the potential to inflict social and psychological harm on the moderators). However, at the same time, AI is not a silver bullet and we identify **10 issues with the use of AI for detecting hateful content**. We argue that AI should supplement rather than supplant the use of human moderators to create effective moderation systems.
3) **Handle online hate** by taking appropriate responses to it: We identify **14 moderation interventions** open to VSPs, each of which imposes different levels of *friction*. We split them into four buckets: a) Hosting constraints b) Viewing constraints c) Searching constraints and d) Engagement constraints. Imposing friction is not costless, however, and concerns have been raised about the implications for freedom of expression, especially for content that is wrongly identified as hateful. The degree of friction which is applied should align with the degree of harm which is likely to be inflicted.

4) **Enable user complaints through a robust and accessible review procedure:** All content moderation systems will make mistakes and so it is vital that users can appeal the decisions they are subjected to. Transparency is key. Important considerations include how much information users are given, whether (and how) users are involved in the content moderation process and the speed with which users' content is moderated.

Beyond content moderation, we examine **other strategies to counter online hate**, including **media literacy** and **counterspeech**. We argue that all efforts to tackle online hate need to be considered in relation to other significant ethical and social concerns such as **freedom of expression, privacy and fairness**, and we elaborate **six risks of excessive moderation of online hate**.

In our **recommendations** we outline six considerations for tackling online hate, based on our analyses in this report.

Recommendations

Based on our analysis and the evidence we have gathered in this report, we present six recommendations to help inform platforms' work to tackle online hate.

1. **Develop a clear account of online hate internally:** A clear characterisation of online hate, with as much clarity and detail as possible, is important for ensuring consistency and fairness in how online hate is tackled. Such an account should engage with the particularly complex issues in online hate, such as self-hatred; truth and validity; and humour and irony.

2. **Match interventions to the severity of hate:** Different types of online hate inflict different degrees and types of harm. Different interventions should be used depending on the harm of the content and the intervention's feasibility (which is related to how resource-intensive and technically complex is the intervention).

3. **Document and explain how online hate is identified:** Identifying online hate is difficult, especially for covert varieties and rarer forms of hate. The processes used to identify online hate (e.g., user reports, AI and human reviews) need to be documented, outlining how they are created, maintained and evaluated, as well as how effective they are.

4. **Explicitly consider the ethical issues in moderating online hate:** Tackling online hate raises fundamental ethical questions, particularly the complex balancing act between protecting users from harm whilst ensuring others' right to freedom of expression is protected. Tackling online hate should not be at the expense of other important concerns.

5. **Policies should be communicated to stakeholders in an understandable way**: It is essential that platforms are transparent, and that all stakeholders are informed

of relevant policies and updates in an accessible and digestible way. Different stakeholders may require different information (i.e., users may only want to know the broad principles of hate moderation rather than the level of detail required by regulators).

6. **Consider approaches beyond content moderation**: Content moderation is the main way in which most platforms tackle online hate. It is a tractable and effective solution. Nonetheless, it is vital that other approaches are considered and that an evidence base is developed to tackle the root causes of online hate.

Finally, we caution that tackling online hate is complex and contentious work, and it often attracts controversy and opposition. Yet it is crucial for ensuring that victims are not left unprotected and that all are able to enjoy the opportunities and affordances created by online platforms. As we wrote in our previous report, *An Agenda for Research Into Online Hate*, online hate is a 'wicked' problem: "it is difficult to define, knowledge is incomplete and contradictory, solutions are not straightforwardly 'good' or 'bad', and it is interconnected with many other problems in society."[5] Ultimately, online hate will not be 'solved' by any one piece of regulation but will require continued and sustained engagement from a range of stakeholders, including the targets of hate and their communities.

[5] Bertie Vidgen et al., *An agenda for research into online hate* (London: The Alan Turing Institute, 2020). Available at: https://www.turing.ac.uk/research/publications/agenda-research-online-hate.

About

The research team

Dr Bertie Vidgen is a Research Fellow in Online Harms within the Public Policy Programme at The Alan Turing Institute, a Research Associate at the University of Oxford and Visiting Fellow at the Open University.[6] He is owner of the hate speech task on the adversarial machine learning platform *Dynabench*, organiser of *The Workshop on Online Abuse and Harms* (ACL), co-Investigator of the *Detecting Online Harms* project and Research Lead on *Hate Speech: Measures & Counter-Measures* (both at the Turing). He co-founded and maintains the website http://ckan.hatespeechdata.com and in 2021 is guest editing two special issues of journals on online harms. Previously, he completed his PhD at the University of Oxford, where he researched online hate.

Emily Burden is a Research Assistant at The Alan Turing Institute and Doctoral Student in Web Science at the University of Southampton.[7]

Professor Helen Margetts OBE 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. She is a Fellow of the British Academy.[8]

The authors thank Josh Cowls and Paul Röttger for their invaluable research assistance and advice during the writing of this report.

[6] Bertie Vidgen, profile on The Alan Turing Institute website. Available at: https://www.turing.ac.uk/people/researchers/bertie-vidgen. Last accessed on 4 December 2020.

[7] Emily Burden, profile on Southampton University's website. Available at: https://www.southampton.ac.uk/history/postgraduate/research_students/elb1g13.page. Last accessed on 15 February 2021.

[8] Helen Margetts, profile on The Alan Turing Institute website. Available at: https://www.turing.ac.uk/people/programme-directors/helen-margetts. Last accessed on 4 December 2020.

The Alan Turing Institute

The Alan Turing Institute is the UK's national institute for data science and artificial intelligence. It was founded in 2015 and undertakes research to tackle some of the biggest challenges in science, society and the economy. It collaborates with universities, businesses and public and third sector organisations, including its 13 partner universities. The Turing aims to help to make the UK the best place in the world for data science and AI research, collaboration, and business.

Hate Speech: Measures and Counter-Measures is part of The Alan Turing Institute's Public Policy Programme. It develops and applies advanced computational methods to measure, analyse and counter hate speech across different online domains.[9] Previously published policy briefing papers include *How Much Online Abuse Is There? A systematic review for the UK*[10] and *An Agenda for Research Into Online Hate*.[11] The project is funded by the Criminal Justice theme within the AI for Science and Government programme under the EPSRC Grant EP/T001569/1.

Funding statement

This report was commissioned and funded by Ofcom to support the implementation of their new VSP regulatory duties.

[9] Hate Speech: Measures and Counter Measures, project page on The Alan Turing Institute website. Available at: https://www.turing.ac.uk/research/research-projects/hate-speech-measures-and-counter-measures. Last accessed on 4 December 2020.

[10] Bertie Vidgen et al., *How Much Online Abuse Is There? A Systematic Review of Evidence for the* UK (London: The Alan Turing Institute, 2019). Available at: https://www.turing.ac.uk/sites/default/files/2019-11/online_abuse_prevalence_full_24.11.2019_-_formatted_0.pdf

[11] Bertie Vidgen et al., *An Agenda for Research Into Online Hate*.

Table of contents

Table of references

Term	Acronym
AVMSD	Audiovisual Media Services Directive
VSP	Video Sharing Platform
NetzDG	Network Enforcement Act
VoD	Video on Demand
AI	Artificial Intelligence
CPS	Crown Prosecution Service

List of Figures, Tables and Boxes

Part 1: Policy-making to tackle online hate

Numerous events have shown the harmful impact of online hate, including the Christchurch massacre in New Zealand in March 2019,[12] the #BlackLivesMatter protests in summer of 2020 and the Capitol riots in Washington in January 2021.[13] At the same time, civil society organisations, such as the Centre for Countering Digital Hate[14] and Glitch![15] in the UK, have campaigned for platforms and governments to take greater action to tackle online hate. Groups such as Stop Hate For Profit[16] and Sleeping Giants[17] have also been successful in defunding advertising revenue from online platforms due to a perceived lack of action against online hate. Pressure has increased from 'above' as well, with more recognition from both governments and international government organisations of the importance of challenging online hate. In 2019's Strategy and Plan of Action on Hate Speech the United Nations Secretary-General António Guterres announced the need to address "the misuse of the Internet and social media for spreading hate speech" (p. 4).[18]

[12] Natasha Tusikov, "Defunding Hate: PayPal's Regulation of Hate Groups." *Surveillance & Society* 17, no. 1/2, pp. 46-53 (2019). Available at: https://ojs.library.queensu.ca/index.php/surveillance-and-society/article/view/12908/8475.

[13] BBC News, "Capitol riots: How a Trump rally turned deadly", 7 January 2020. Available at: https://www.bbc.co.uk/news/av/world-us-canada-55569495.

[14] The Centre for Countering Digital Hate. Available at: https://www.counter-hate.com. Last accessed on 4 December 2020.

[15] Glitch!, https://fixtheglitch.org/online-abuse/. Last accessed on 4 December 2020.

[16] Stop Hate for Profit, https://www.stophateforprofit.org/. Last accessed on 4 December 2020.

[17] Sleeping Giants, https://www.slpnggiants.com/. Last accessed on 4 December 2020.

[18] United Nations, *United Nations Strategy and Plan of Action on Hate Speech* (New York: United Nations, 2019). Available at: https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf.

Since 2018 many countries have implemented new regulations. In Germany, the Network Enforcement Act (NetzDG)[19] came into effect on January 1st 2018 to combat online hate speech. Under this law, online platforms face a maximum fine of 50 million euros if they fail to remove illegal content. NetzDG initially attracted criticism from those, such as Germany's best-selling newspaper Bild, who viewed it as part of an overly punitive censorship regime which would 'chill' free speech.[20] However, NetzDG has resulted in fewer takedowns than some originally feared. As The Counter-Extremism Project summarises, "NetzDG has not provoked mass requests for takedowns. Nor has it forced online platforms to adopt a "take down, ask later' approach" (p. i).[21]

In Australia, legislation was passed in 2019 that criminalises the sharing of "abhorrent violent material."[22] The law was passed a month after the Christchurch terrorist attack in New Zealand and establishes fines for platforms if they do not remove content "expeditiously". The employees of platforms could also be sentenced to up to three years in prison.[23] In France, the anti-online hate "Avia" law was passed by the National Assembly in July 2019 with support from President Macron's government. This legislation obligated platforms to remove flagged hateful and extremist content within 24

[19] German Federal Ministry of Justice and Consumer Protection, *Act To Improve Enforcement Of The Law In Social Networks (Network Enforcement Act)* (Berlin: Germany, 2017). Available at: https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.html.

[20] Philip Oltermann, "Tough new German law puts tech firms and free speech in the spotlight", *The Guardian,*
5 January 2018. Available at: https://www.theguardian.com/world/2018/jan/05/tough-new-german-law-puts-tech-firms-and-free-speech-in-spotlight.

[21] William Echikson and Olivia Knodt, *Germany's NetzDG: A key test for combating online hate* (Brussels: Counter-Extremism Project, 2018). Available at: http://wp.ceps.eu/wp-content/uploads/2018/11/RR%20No2018-09_Germany's%20NetzDG.pdf.

[22] Australia Legislation, *Criminal Code Amendment (Sharing of Abhorrent Violent Material) Bill 2019* (Canberra: Australia, 2019). Available at: https://www.legislation.gov.au/Details/C2019A00038.

[23] Damien Cave, "Australia Passes Law to Punish Social Media Companies for Violent Posts", *New York Times,* 3 April 2019. Available at: https://www.nytimes.com/2019/04/03/world/australia/social-media-law.html.

hours or risk a fine of up to 1.25 million euros.[24] However, in June 2020 the law was heavily amended by the Constitutional Council who stated that otherwise it would "infringe upon the exercise of freedom of expression and communication in a way that is not necessary, suitable, and proportionate."[25]

1.1 Recent policy-making developments in the UK

In the UK, the Online Harms White Paper was released by the Department for Digital, Media, Culture and Sport and the Home Office in April 2019. It announced the need to clean up, regulate and monitor online spaces for myriad harms. [26] In its response to the Online Harms White Paper consultation, published in December 2020, the UK Government reiterated its commitment to addressing the spread and impact of harmful online content:

> "The COVID-19 pandemic has shone a spotlight on the risks posed by harmful activity and content online. The pandemic drove a spike in disinformation and misinformation, and some people took advantage of the uncertainty to incite fear and cause confusion."[27]

[24] Aurelien Breeden, "French Court Strikes Down Most of Online Hate Speech Law", *New York Times*, 18 June 2020, https://www.nytimes.com/2020/06/18/world/europe/france-internet-hate-speech-regulation.html.

[25] French Constitutional Council Press Release, "Decision 2020-801 DC of June 18, 2020 press release", (Paris: French Constitutional Council, 18 June 2020). Available at: https://www.conseil-constitutionnel.fr/actualites/communique/decision-n-2020-801-dc-du-18-juin-2020-communique-de-presse.

[26] Department for Digital, Culture, Media & Sport and the Home Office, *Online Harms White Paper* (London: UK Government, 2019). Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf.

[27] Department for Digital, Culture, Media & Sport and the Home Office, *Online Harms White Paper: Full Government Response to the consultation*.(London: UK Government, 2020). Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/944310/Online_Harms_White_Paper_Full_Government_Response_to_the_consultation_CP_354_CCS001_CCS1220695430-001__V2.pdf.

Notably, the Online Harms White Paper outlines the need to protect individuals from both illegal content and content that is "harmful but legal".[28] This extra-legal category has generated considerable debate as to what regulatory mechanisms should be used to address it, as well as how protecting people from legal but harmful content should be balanced with other concerns around privacy and freedom of expression.[29] It is worth noting that the need to address harmful but legal hateful content has also been proposed by other bodies in the UK. The Commission for Countering Extremism, a Home Office supported organisation to fight all forms of extremism in the UK, notes that "extremist groups can engage in hateful behaviours directed at minority groups, which is not illegal or criminal"[30] and the Law Commission acknowledges that online hate may not always be a hate crime: "By "online hate" we mean a hostile online communication that targets someone on the basis of an aspect of their identity (including but not limited to protected characteristics). Such communications will not necessarily amount to a hate crime." (p.197)[31] As of December 2020, the Government has announced that it expects new

[28] Ibid.

[29] Institute for Strategic Dialogue, *A joint statement on the Online Harms White Paper and the direction of regulation in the UK* (London: Institute for Strategic Dialogue, 2020). Available at: https://www.isdglobal.org/isd-publications/joint-statement-on-the-online-harms-white-paper/.

[30] Commission for Countering Extremism, *Challenging Hateful Extremism* (London: UK Home Office, 2019). Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/874101/200320_Challenging_Hateful_Extremism.pdf.

[31] Law Commission, *Harmful Online Communications: The Criminal Offences - A Consultation Paper* (London: Law Commission, 2020). Available at: https://s3-eu-west-2.amazonaws.com/lawcom-prod-storage-11jsxou24uy7q/uploads/2020/09/Online-Communications-Consultation-Paper-FINAL-with-cover.pdf.

legislation for an Online Harms Bill to be introduced in 2021[32] and has confirmed appointment of Ofcom as the new online harms regulator.[33]

The new VSP Framework is part of an evolving and interrelated landscape of online regulations in the UK, and internationally. The UK Government is in the process of bringing forward online safety legislation which will supersede the narrower VSP Framework brought in under the AVMSD in November 2020. Since 2018 the Law Commission, the UK's statutory body for reviewing and updating the law, has launched two consultations which could lead to changes in the laws relating to online hate. One consultation is on offensive communications[34] and the other is on hate crime.[35] In principle, online and offline hate are given equal levels of protection under UK law; in 2017, the Director of Public Prosecutions at the Crown Prosecution Service (CPS) announced that the CPS "commits to treat online hate crimes as seriously as those committed face to face"[36] and a report in that same year by the Home Affairs Select

[32] UK Safer Internet Centre, "Government says new online harms legislation is expected to be ready next year" (London: UK Safer Internet Centre, 9 October 2020), Available at: https://www.saferinternet.org.uk/blog/government-says-new-online-harms-legislation-expected-be-ready-next-year.

[33] UK Government Press Release, "Government minded to appoint Ofcom as online harms regulator", 12 February 2020. Available at: https://www.gov.uk/government/news/government-minded-to-appoint-ofcom-as-online-harms-regulator;
Ofcom, "Ofcom to regulate harmful content online", 15 December 2020. Available at: https://www.ofcom.org.uk/about-ofcom/latest/features-and-news/ofcom-to-regulate-harmful-content-online.

[34] Law Commission, "Consultation on the reform of the communications offences", 24 September to 18 December 2020. Available at: https://consult.justice.gov.uk/law-commission/online_comms/.

[35] Law Commission, "Hate Crime Consultation", https://www.lawcom.gov.uk/project/hate-crime/. Last accessed on 4 December 2020.

[36] Alison Saunders, "Hate is hate. Online abusers must be dealt with harshly," *The Guardian,* 21 August 2017. Available at: https://www.theguardian.com/commentisfree/2017/aug/20/hate-crimes-online-abusers-prosecutors-serious-crackdown-internet-face-to-face.

Committee stated, “The Government has been clear that what is illegal offline is also illegal online in relation to hate speech and abuse.” (p. 18)[37]

Two main protections are provided against hate. First, cases where an existing criminal activity (such as property damage or physical violence) is shown to be motivated by prejudice and the offence is “aggravated” and receives enhanced sentencing.[38] Second, cases where hateful language is used. In most offline contexts, hate has generally been prosecuted by the CPS under Part III of the 1986 Public Order Act, which prohibits “acts intended or likely to stir up racial hatred” against “a group of persons defined by reference to colour, race, nationality (including citizenship) or ethnic or national origins.” The legislation stipulates:

> “A person who uses threatening, abusive or insulting words or behaviour, or displays any written material which is threatening, abusive or insulting, is guilty of an offence if—
>
> (a) he intends thereby to stir up racial hatred, or
> (b) having regard to all the circumstances racial hatred is likely to be stirred up thereby.”[39]

In its 2020 consultation paper on *Harmful Online Communications* the Law Commission noted that online hate may also be prosecuted under the communications offences.[40] Section 1 of the Malicious Communications Act 1988 (MCA 1988) and section 127 of the

[37] Home Affairs Committee*, Hate crime: abuse, hate and extremism online - Fourteenth Report of Session 2016–17* (London: UK Government, 2017). Available at: https://publications.parliament.uk/pa/cm201617/cmselect/cmhaff/609/609.pdf.

[38] Crown Prosecution Service, *Hate Crime Report 2018-2019* (London: Crown Prosecution Service, 2019). Available at: https://www.cps.gov.uk/sites/default/files/documents/publications/CPS-Hate-Crime-Annual-Report-2018-2019.PDF.

[39] UK Legislation, *Public Order Act 1986* (London: UK,1986). Available at: https://www.legislation.gov.uk/ukpga/1986/64.

[40] Law Commission, *Harmful Online Communications: The Criminal Offences - A Consultation Paper.*

Communications Act 2003 (CA 2003) contain the most relevant provisions.[41] Section 127 addresses “improper use” of public electronic communications networks. It states that a person is guilty of an offence if s/he “(a) sends by means of a public electronic communications network a message or other matter that is grossly offensive or of an indecent, obscene or menacing character; or (b) causes any such message or matter to be so sent.” The provision for “grossly offensive” communications, such as those with a “menacing” character, have been used as the legal basis for prosecution of online hate. In January 2017 Rhodri Philipps, the 4th Viscount St Davids, posted on Facebook that he would pay "£5,000 for the first person to 'accidentally' run over this bloody troublesome first generation immigrant" about anti-Brexit campaigner Gina Miller and "If this is what we should expect from immigrants, send them back to their stinking jungles".[42] Phillips was found guilty under the Communications Act and charged with “malicious communications with racially aggravated factors”.[43]

Box 1: The limits of the law for tackling online hate

In 2012, Port Talbot football player Daniel Thomas sent a homophobic tweet referencing Olympic divers Tom Daley and Peter Waterfield: “if there is any consolation for finishing fourth at least daley and waterfield can go and bum each other #teamHIV”.[44] The tweet was posted publicly and did not use an @ mention to address

[41] Law Commission, *Hate Crime: Consultation Paper Summary* (London: Law Commission, 2020). Available at: https://s3-eu-west-2.amazonaws.com/lawcom-prod-storage-11jsxou24uy7q/uploads/2020/09/Hate-crime-final-summary.pdf.

[42] BBC News, “Aristocrat guilty over 'menacing' Gina Miller Facebook post”, July 11 2017. Available at: https://www.bbc.co.uk/news/uk-40574754.

[43] Kevin Rawlinson, “Viscount who was jailed over Gina Miller threats drops his appeal”, *The Guardian*, 25 August 2017. Available at: https://www.theguardian.com/politics/2017/aug/25/viscount-jailed-gina-miller-threats-drops-appeal-sentence.

[44] Martin Hickman, "Chief prosecutor reveals lenient stance after footballer is cleared of abusing Tom Daley", *The Independent*, 20 September 2012. Available at: https://www.independent.co.uk/news/uk/crime/chief-prosecutor-reveals-lenient-stance-after-footballer-cleared-abusing-tom-daley-8160648.html.

it directly to Daley or Waterfield. Thomas was arrested for sending a malicious communication and referred to the CPS to consider whether he should be charged with a criminal offence. Following consultation with Daley and Waterfield, the Director of Public Prosecutions determined that the communication fell below the threshold for criminal prosecution. This was based on a verdict that the tweet was not intended to reach Daley and Waterfield, was not "grossly offensive", and was not part of a campaign. Thomas also showed remorse for causing offence and removed the message.[45] The Director of Public Prosecutions concluded, "The fact that offensive remarks may not warrant a full criminal prosecution does not necessarily mean that no action should be taken."[46]

1.2 Policy-making at the EU-level

Policies to tackle online hate have been developed by the European Union (EU). In May 2016, following three coordinated terrorist attacks in Brussels, the EU Code of Conduct to Tackle Online Hate was launched by the European Commission.[47] It started with four major IT companies (Facebook, Microsoft, Twitter and YouTube), and aimed to respond to the proliferation of terrorist content and racist and xenophobic hate speech online. In 2018, Instagram, Snapchat and Dailymotion joined the Code of Conduct, Jeuxvideo.com joined in January 2019 and TikTok in September 2020. By joining companies commit to review "the majority of [takedown] requests in less than 24 hours and to removing the content if necessary, while respecting the fundamental principle of freedom of speech."

[45] BBC News, "Tom Daley Tweet: No Action Against Daniel Thomas", 20 September 2012. Available at: https://www.bbc.co.uk/news/uk-wales-19661950.

[46] BBC News, "Tom Daley 'Abuse' Tweet: Legal Rethink On Online Rules", 20 September 2012. Available at: https://www.bbc.co.uk/news/uk-19660415.

[47] Council of the European Union Press Release, "Joint statement of EU Ministers for Justice and Home Affairs and representatives of EU institutions on the terrorist attacks in Brussels on 22 March 2016", 24 March 2016. Available at: https://www.consilium.europa.eu/en/press/press-releases/2016/03/24/statement-on-terrorist-attacks-in-brussels-on-22-march/.

The Code's implementation has been evaluated through regular monitoring exercises whereby trusted civil society organisations notify the platforms of content they deem to be illegal hate, and then monitor how they respond. In the UK, the trusted organisations are the Media Diversity Institute, Galop, Community Security Trust and Tell Mama.[48] The fifth evaluation was conducted in June 2020.[49] It found that across all platforms 90% of notifications were reviewed within 24 hours and 71% of the flagged content was removed. These figures are broadly stable with the previous evaluation and reflect a marked improvement on the first evaluation (in 2016), where only 40% of flagged content was reviewed within 24 hours and 28.2% was removed. The report for the fifth evaluation reported that more work is still needed to tackle online hate as "some divergences exist among the platforms. Most of the IT companies must improve their feedback to users' notifications."[50] However, it also noted that a 100% takedown rate is undesirable given that it could mean platforms are being too draconian, which risks penalising legitimate free speech.

Other work has been conducted at the European level to address online hate. At the request of The European Parliament Committee on Internal Market and Consumer Protection, a study was conducted in June 2020 to review the EU regulatory approach to content moderation and the practices of online platforms. It made recommendations to improve the EU legal framework within the context of the forthcoming Digital Services Act.[51] The report noted:

[48] Didier Reynders, *Countering Illegal Hate Speech Online - 5th Evaluation of the Code of Conduct* (Brussels: European Commission, 2020). Available at: https://ec.europa.eu/info/sites/info/files/codeofconduct_2020_factsheet_12.pdf.

[49] The first monitoring exercise was completed in December 2016; the second in May 2017; the third in December 2018; the fourth in December 2018; and the fifth in June 2020.

[50] Didier Reynders, *Countering Illegal Hate Speech Online - 5th Evaluation of the Code of Conduct*.

[51] Alexandre de Streel et al., *Online Platforms' Moderation of Illegal Content Online: Law, Practices and Options for Reform* (Luxembourg: Policy Department for Economic, Scientific and Quality of Life Policies, 2020). Available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/652718/IPOL_STU(2020)652718_EN.pdf
.

> “...the Counter-Racism Framework Decision provides that Member States must ensure that racist and xenophobic hate speech is punishable, **but does not impose detailed obligations related to online content moderation practice** [...] In addition to this multi-layered EU regulatory framework, several Member States have adopted national rules on online content moderation in particular for hate speech and online disinformation. **The legal compatibility of those national initiatives with the EU legal framework is not always clear** and the multiplication of national laws seriously risks undermining the Digital Single Market” (p.10, emphasis added)

1.3 The Audiovisual Media Services Directive (AVMSD)

The AVMSD governs EU-wide coordination of national legislation on all audiovisual media, both traditional TV broadcasts, ‘video on demand’ (VOD) services and VSPs.[52] It originated from the Television without Frontiers Directive (89/552/EEC)[53] which was adopted in 1989 to ensure the free movement of broadcasting services within the internal market and at the same time to preserve certain public interest objectives, such as cultural diversity, the right of reply, consumer protection and the protection of minors.[54] A revised directive was adopted in 1997 to establish the ‘country of origin’ principle and update the initial rules, for example to place greater emphasis on the protection of minors. The Directive was renamed the AVMSD in 2007 and further revised to account for VOD services which were becoming increasingly available via the Internet. The Directive was consolidated into Directive 2010/13 in 2010[55] and amended in 2018 by

[52] UK Legislation, *The Audiovisual Media Services Regulations 2020.*

[53] EU Legislation, *Television broadcasting activities: "Television without Frontiers" (TVWF) Directive* (Brussels: European Union, 1989). Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=LEGISSUM%3Al24101.

[54] European Parliament, “Audiovisual and media policy”. Available at: https://www.europarl.europa.eu/factsheets/en/sheet/138/audiovisual-and-media-policy. Last accessed on 4 December 2020.

[55] EU Legislation, *Directive 2010/13/EU of the European Parliament and of the Council of 10 March 2010* (Brussels: European Union, 2010). Available at: http://data.europa.eu/eli/dir/2010/13/oj.

Directive 2018/1088 to account for “new types of services and user experiences”, in particular Video Sharing Platforms (VSPs).[56] This change had been proposed in May 2016 by the European Commission as part of its Digital Single Market Strategy[57] following a public consultation in 2015 to understand “how to make Europe's audiovisual media landscape fit for purpose in the digital age”.[58]

A key purpose of the 2018 Directive was to introduce a new regulatory framework for VSPs, requiring providers of such services to take appropriate measures to “protect children (under 18s) from content which might impair their physical, mental or moral development” and to protect the general public from “content inciting violence or hatred, and content constituting criminal offences relating to terrorism; child sexual exploitation and abuse and child pornography; and racism and xenophobia.” (p. 1)[59] Under the AVMSD a VSP is defined as “a service or dissociable section of a service [...] where the provision of videos to members of the public is (a) the principal purpose of the service or of the dissociable section of the service, or (b) an essential functionality of the service.” This means that platforms which are not primarily designed to enable video sharing may still fall under the remit of the AVMSD.[60]

[56] EU Legislation Proposal, *Proposal for a Directive of the European Parliament and of the Council amending Directive 2010/13/EU* (Brussels: European Union, 2016). Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1464618463840&uri=COM:2016:287:FIN.

[57] European Commission Press Release, “A Digital Single Market for Europe: Commission sets out 16 initiatives to make it happen”, 6 May 2015, https://ec.europa.eu/commission/presscorner/detail/en/IP_15_4919.

[58] European Commission, “Public consultation on Directive 2010/13/EU on Audiovisual Media Services (AVMSD) - A media framework for the 21st century”, 6 July 2015 to 30 September 2015. Available at: https://ec.europa.eu/digital-single-market/en/news/public-consultation-directive-201013eu-audiovisual-media-services-avmsd-media-framework-21st.

[59] Ofcom, *Regulating video-sharing platforms A guide to the new requirements on VSPs and Ofcom’s approach to regulation* (London: Ofcom, 2020). Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0021/205167/regulating-vsp-guide.pdf.

[60] EU Legislation, *Directive (EU) 2018/1808 of the European Parliament and of the Council of 14 November 2018 amending Directive 2010/13/EU* (Brussels: European Union, 2018). Available at: http://data.europa.eu/eli/dir/2018/1808/oj.

The AVMSD outlines the importance of VSPs taking steps to tackle online hate. Recital 47 of the Directive refers to VSPs taking appropriate measures:

> "to protect the general public from content that contains incitement to violence or hatred directed against a group or a member of a group on any of the grounds referred to in Article 21 of the Charter of Fundamental Rights of the European Union (the 'Charter'), or the dissemination of which constitutes a criminal offence under Union law." (emphasis added)[61]

Article 21 of the EU's Charter of Fundamental Rights outlines 14 facets of identity which could be the basis of hatred:

> "Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited." (emphasis added)[62]

Recital 17 of the 2018 AVMSD explains "the notion of 'incitement to violence or hatred" should be "understood within the meaning of Council Framework Decision 2008/913/JHA". This Decision describes what racist and xenophobic conduct is punishable under EU law.[63] Paragraph 1 of Article 1 of the Decision contains four sub-paragraphs. Sub-paragraph (a) prohibits "publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin." Sub-paragraph (b) prohibits "the commission of an act referred to in sub-paragraph (a) by public dissemination or

[61] Ibid.

[62] EU Legislation, *EU Charter of Fundamental Rights 2012* (Brussels: European Union, 2012). Available at: https://fra.europa.eu/en/eu-charter/article/21-non-discrimination.

[63] EU Legislation, *Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law* (Brussels: European Union, 2008). Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2008.328.01.0055.01.ENG&toc=OJ:L:2008:328:TOC.

distribution of tracts, pictures or other material”. Sub-paragraph (c) criminalises “publicly condoning, denying or grossly trivialising crimes of genocide, crimes against humanity and war crimes as defined in Articles 6, 7 and 8 of the Statute of the International Criminal Court, directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin when the conduct is carried out in a manner likely to incite to violence or hatred against such a group or a member of such a group”. Sub-paragraph (d) extends the behaviour proscribed in (c) to “crimes defined in Article 6 of the Charter of the International Military Tribunal”.

Paragraphs 2 and 3 set out further additional detail in relation to the offences in paragraph 1. Paragraph 2 outlines that Member States have discretion to decide that conduct will only be prosecuted if it disrupts public behaviour:

> “For the purpose of paragraph 1, Member States may choose to punish only conduct which is either carried out in a manner likely to disturb public order or which is threatening, abusive or insulting.”

Paragraph 3 explains:

> “Each Member State shall take the measures necessary to ensure that instigating the conduct referred to in Article 1(1)(c) and (d) is punishable.
>
> Each Member State shall take the measures necessary to ensure that aiding and abetting in the commission of the conduct referred to in Article 1 is punishable.”

Finally, Article 4 provides measures for aggravated sentencing where other criminal conduct has “racist and xenophobic motivation”.

1.4 Minors and online hate

This report does not address minors as a special focus, although we recognise that online hate can impact them as well as adults. One concern is that minors will be exposed to hateful content through the malicious activities of hateful individuals and groups. This is a plausible risk, shown by the co-optation of a My Little Pony fan site, Derpibooru, by white supremacists in 2020.[64] An investigation by The Atlantic found evidence of racist and violent content on the fan forum (which is primarily aimed at children), with more than 900 pieces of art explicitly tagged as such. Further, the COVID-19 pandemic could open new avenues through which children could be impacted by online hate.[65] Restrictions imposed on school, socialising and work has meant children are spending more time online for educational, entertainment and communication purposes. In turn, they are being exposed to increased levels of harmful content, with potentially negative impact on their mental health and development.[66] Research by the British Board of Film Classification in May 2020 found that 47% of teens in the survey said they had seen content online they wish they had not, and one in seven (13%) said they see harmful videos every day.[67]

Some strategies to specifically support minors and to make them more aware of the harm inflicted by online hate have already been developed. For instance, the BBC has created the *Own It App* to support children as they navigate online spaces through their mobile

[64] Kaitlyn Tiffany, "My Little Pony Fans Are Ready to Admit They Have a Nazi Problem", *The Atlantic*, 23 June 2020. Available at: https://www.theatlantic.com/technology/archive/2020/06/my-little-pony-nazi-4chan-black-lives-matter/613348/.

[65] Robyn Millar et al., *Considering the evidence of the impacts of lockdown on the mental health and wellbeing of children and young people within the context of the individual, the family, and education* (Glasgow: Mental Health Foundation, 2020). Available at: https://www.mentalhealth.org.uk/sites/default/files/MHF%20Scotland%20Impacts%20of%20Lockdown.pdf.

[66] Pouria Babvey et al., "Using Social Media Data for Assessing Children's Exposure to Violence during the COVID-19 Pandemic", *Child Abuse & Neglect* [in proof], (2020). Available at: https://doi.org/10.1016/j.chiabu.2020.104747;

Joanne Orlando, "Young people are exposed to more hate online during COVID. And it risks their health", *The Conversation*, 9 November 2020. Available at: https://theconversation.com/young-people-are-exposed-to-more-hate-online-during-covid-and-it-risks-their-health-148107.

[67] British Board of Classification, "Half of children and teens exposed to harmful online content while in lockdown", 4 May 2020. Available at: https://www.bbfc.co.uk/about-us/news/half-of-children-and-teens-exposed-to-harmful-online-content-while-in-lockdown.

phones.[68] The *Own It App*'s flagship offering is a custom keyboard which becomes the default for all text input fields on the child's mobile phone – including all messaging apps and web pages. In order to help minors determine the potential impact of their online behaviour it provides live feedback about the sentiment of messages before they are sent. This tool is powered by technology which assesses messages for hate, toxicity, safeguarding and privacy, as well as other aspects.

[68] BBC, "Own It, The App: Six Technical Challenges", 18 September 2019. Available at: https://www.bbc.co.uk/blogs/internet/entries/94ec41ae-b25b-4e58-9c0f-1b9b2890c281.

Part 2: Understanding online hate

Online hate is a deeply contested and complex concept. To guide this report, we adopt the following definition, taken from forthcoming work by The Alan Turing Institute:[69]

> "Online hate speech is a communication on the Internet which expresses prejudice against an identity. It can take the form of derogatory, demonising and dehumanising statements, threats, identity-based insults, pejorative terms and slurs. Online hate speech involves:
>
> 1. A medium for the content such as text, images, video, audio, and gifs;
> 2. A perpetrator (the person who creates or shares the hateful content);
> 3. An actual or potential audience (anyone who is or who could be exposed to or targeted by the content);
> 4. A communicative setting (e.g., private messaging apps, online forums, comment sections or broadcast-style social media platforms)."

2.1 How online and offline hate differ

Online and offline hate are, in principle, treated equivalently under UK law and both are widely recognised as being equally capable of inflicting harm on victims. As David Kaye, the UN Special Rapporteur on the promotion and protection of the right to freedom of expression and opinion, put it in 2019: "[o]nline hate is no less harmful because it is online".[70] However, there are key differences between online and offline hate in terms of their nature, dynamics and prevalence. A report by UNESCO identified that "while hate speech online is not intrinsically different from similar expressions found offline" it has distinct features, such as the fact that it is typically permanent (as online content is

[69] Based on forthcoming work from The Alan Turing Institute, developed by Bertie Vidgen, Josh Cowls and Helen Margetts.

[70] UN Office of the High Commissioner for Human Rights, "Governments And Internet Companies Fail To Meet Challenges Of Online Hate – UN Expert", 21 October 2019. Available at: https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=25174&LangID=E.

usually hosted indefinitely), can easily ‘travel’ across the web to reach large and varied audiences, is often created by people who are anonymous, and its production can cross multiple legal jurisdictions (p. 13).[71] Similarly, in a paper titled, *What is so special about online (as compared to offline) hate speech?* legal scholar Brown outlines several distinctive features of online hate, including (1) the ease with which purveyors of hate can access audiences, (2) the size of the audiences they reach, (3) their anonymity and (4) the instantaneousness of sending hate.

Of all the distinctive features of online hate, anonymity has arguably received the greatest attention. In an early paper on online hate regulation, Citron and Norton contend that the anonymous and pseudonymous nature of online discourse “can just as easily accelerate destructive behaviour as it can fuel public discourse”.[72] Similarly Boyd, from The Sentinel Project has stated, “people feel much more comfortable speaking hate online as opposed to real life when they have to deal with the consequences of what they say”.[73] This is the so-called ‘disinhibition effect’: individuals may be more aggressive and hateful online because they cannot see how victims are affected by their content.[74] Brown proposes that disinhibition can also occur because the Internet enables individuals to engage in more spontaneous and unconsidered communications, with little time spent considering how their behaviour might impact victims.[75]

Online hate can be expressed through many types of media, including text, images, videos and audio. In some media hate is expressed multimodally, in which different types of communication are combined. This can allow for more complex forms of expression

[71] Iginio Gagliardone et al., *Countering online hate speech* (Paris: UNESCO, 2017). Available at: https://unesdoc.unesco.org/ark:/48223/pf0000233231.

[72] Danielle Citron and Helen Norton, “Intermediaries and hate speech: fostering digital citizenship for our information age”, *Boston University Law Review*, 91:16, pp. 1435-1484 (2011). Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1764004.

[73] Iginio Gagliardone et al., *Countering online hate speech*.

[74] Alexander Brown, "What Is So Special About Online (As Compared to Offline) Hate Speech?", *Ethnicities*, 18: 13, pp. 297-326 (2017). Available at: https://doi.org/10.1177/1468796817709846.

[75] Ibid.

and is a hallmark of online content. Memes are a multi-modal form of communication which combine an image with text. Some suggest that they have played a key role in facilitating the movement of hateful ideas from the margins to the mainstream of society as memes are often engaging, humorous and innocuous – yet can easily contain deeply prejudicial ideas.[76] Memes are particularly challenging to moderate as hate can be expressed through an otherwise benign image and benign text – but which become hateful when considered together.[77] For instance, an image showing Muslims in the UK could be superimposed with the text, 'We've had enough. We should kick them out!'. If the image were changed to a meeting of the UK cabinet or if the text were changed to 'United in prayer' then the meme would no longer be hateful. Videos, snaps and gifs pose similar challenges in that they layer many modes of communication on top of each other, making the content harder to decipher and therefore more difficult to address. Videos often combine text, images and audio all at once, making them particularly difficult to analyse.

2.2 Evidence on online hate

A large and growing body of research, spanning the social, behavioural and natural scientific research fields, has been devoted to analysing the empirical dynamics of online hate. Given the scope of this report, we offer a high-level summary of previous research:

1. **Assessing the prevalence of online hate is a difficult task due to the lack of appropriate data and robust measurement tools**. In a 2019 report on online abuse The Alan Turing Institute summarised, "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

[76] Aaron Winter, "Online Hate: From the Far-Right to the 'Alt-Right' and from the Margins to the Mainstream," in *Online Othering. Palgrave Studies in Cybercrime and Cybersecurity*, Karen Lumsden and Emily Harmer eds. (London: Palgrave Macmillan, 2019). Available at: https://doi.org/10.1007/978-3-030-12633-9_2.

[77] Douwe Kiela et al., "The hateful memes challenge: Detecting hate speech in multimodal memes", *arXiv:2005.04790* (2020). Available at: https://arxiv.org/abs/2005.04790.

provided with the necessary contextual information to fully interpret them" (p. 5).[78] Notably, in the most recent report on hate crime from the Home Office (covering 2019/2020) statistics on online hate crime were not made available. The last time that such figures were reported was in 2017/2018 when "experimental" figures were given for 30 out of 44 police forces. They showed that 1,605 online hate crimes were recorded in England and Wales, around 2% of all hate crimes.[79] In November 2020, Facebook reported for the first time that the percentage of content exposures for hate speech was 0.11%.[80] This means that for every 1,000 times a piece of content is viewed on the platform, one of them will be hateful content.

2. **The prevalence and dynamics of online hate vary across platforms. In general, there is far less overt hate on more mainstream platforms.** Hine et al. measured the prevalence of online hate during 2016 using the HateBase lexicon.[81] They compared hate on three popular forums (or 'boards') on 4chan, showing that 12% of posts in the forum '/pol/' were hateful, 6.3% of posts in '/sp/' and 7.3% of posts in '/int/'. They also analysed a random sample of posts from Twitter, of which 2% were identified as being hateful. This research indicates that, although not all posts are hateful even in the more extreme parts of the Internet, the level of hate is significantly higher in such spaces than in mainstream platforms. Other studies report similar results about the higher prevalence and more extreme nature of

[78] Bertie Vidgen et al., *How much online abuse is there? A systematic review of evidence for the UK.*

[79] Home Office, *Hate Crime, England and Wales 2017/18* (London: Home Office, 2018). Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/748598/hate-crime-1718-hosb2018.pdf.

[80] Facebook, *Community Standards Enforcement Report Q3 2020* (San Francisco: Facebook, 2020). Available at: https://transparency.facebook.com/community-standards-enforcement#hate-speech.

[81] Gabriel Emile Hine et al., "Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and its Effects on the We", in *Proceedings of the Eleventh International AAAI Conference on Web and Social Media* (Montreal: Association for the Advancement of Artificial Intelligence (AAAI), pp. 92-101 (2017). Available at: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15670.

hate on alternative online platforms such as Gab and Reddit.[82] However, making such assessments is difficult due to the speed at which new platforms emerge and attract users, such as TikTok's growth during 2019 and 2020.[83]

3. **Online hate is perpetrated by both lone individuals (such as 'bedroom' trolls) and individuals affiliated with hate groups, such as white supremacists.**[84] Hateful actors can have a range, including inflicting harm on victims, creating social division, forming and reinforcing their own community cohesion, humour and 'shitposting' (the practice of intentionally posting provocative or off-topic content to disrupt an online discussion).

4. **The prevalence and dynamics of online hate vary across platforms. In general, there is far less overt hate on more mainstream platforms.** Hine et al. measured the prevalence of online hate during 2016 using the HateBase lexicon.[85] They

[82] Savvas Zannettou et al., "What Is Gab? A Bastion of Free Speech or an Alt-Right Echo Chamber?" in *Proceedings of the International World Wide Web Conference ACM* (Lyon: International World Wide Web Conferences), pp. 1007-1014 (2018). Available at: https://doi.org/10.1145/3184558.3191531.;
Enrico Mariconti et al., "You Know What to Do": Proactive Detection of YouTube Videos Targeted by Coordinated Hate Attacks", *Proceedings of the ACM on Human-Computer Interaction*, 3: 207 (2019). Available at: https://dl.acm.org/doi/10.1145/3359309.
Tracie Farrell et al., "Exploring Misogyny across the Manosphere in Reddit", in *Proceedings of the 10th ACM Conference on Web Science* (New York: Association for Computing Machinery), pp. 87–96 (2019). Available at: https://doi.org/10.1145/3292522.3326045.

[83] The NYU Dispatch, "Instagram vs TikTok: The Battle Between Social Media Platforms" (New York: The NYU Dispatch, 20 February 2020). Available at: https://wp.nyu.edu/dispatch/2020/02/20/instagram-vs-tiktok-the-battle-between-social-media-platforms/.

[84] REACT, *National qualitative and quantitative report: United Kingdom* (Milan: React No Hate, 2018). Available at: http://www.reactnohate.eu/wp-content/uploads/2019/09/D2.3_REACT_UK-National-Qualitative-and-quantitative-Report-on-the-monitoring-results.pdf.

[85] Gabriel Emile Hine et al., "Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and its Effects on the We", in *Proceedings of the Eleventh International AAAI Conference on Web and Social Media* (Montreal: Association for the Advancement of Artificial Intelligence (AAAI), pp. 92-101 (2017). Available at: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15670.

compared hate on three popular forums (or ‘boards’) on 4chan, showing that 12% of posts in the forum ‘/pol/’ were hateful, 6.3% of posts in ‘/sp/’ and 7.3% of posts in ‘/int/’. They also analysed a random sample of posts from Twitter, of which 2% were identified as being hateful. This research indicates that, although not all posts are hateful even in the more extreme parts of the Internet, the level of hate is significantly higher in such spaces than in mainstream platforms. Other studies report similar results about the higher prevalence and more extreme nature of hate on alternative online platforms such as Gab and Reddit.[86] However, making such assessments is difficult due to the speed at which new platforms emerge and attract users, such as TikTok’s growth during 2019 and 2020.[87]

5. **Online hate spreads through online networks.** Johnson et al. examined hate networks and the “adaptive dynamics” of the global online hate ecology.[88] They argue that the key to understanding this resilience of online hate networks lies in its “global network-of-network” dynamics. They describe how “Interconnected hate clusters form global ‘hate highways’ that—assisted by collective online adaptations—cross social media platforms, sometimes using ‘back doors’ even after being banned, as well as jumping between countries, continents and

[86] Savvas Zannettou et al., “What Is Gab? A Bastion of Free Speech or an Alt-Right Echo Chamber?” in *Proceedings of the International World Wide Web Conference ACM* (Lyon: International World Wide Web Conferences), pp. 1007-1014 (2018). Available at: https://doi.org/10.1145/3184558.3191531.;
Enrico Mariconti et al., “You Know What to Do”: Proactive Detection of YouTube Videos Targeted by Coordinated Hate Attacks”, *Proceedings of the ACM on Human-Computer Interaction*, 3: 207 (2019). Available at: https://dl.acm.org/doi/10.1145/3359309.
Tracie Farrell et al., “Exploring Misogyny across the Manosphere in Reddit”, in *Proceedings of the 10th ACM Conference on Web Science* (New York: Association for Computing Machinery), pp. 87–96 (2019). Available at: https://doi.org/10.1145/3292522.3326045.

[87] The NYU Dispatch, “Instagram vs TikTok: The Battle Between Social Media Platforms” (New York: The NYU Dispatch, 20 February 2020). Available at: https://wp.nyu.edu/dispatch/2020/02/20/instagram-vs-tiktok-the-battle-between-social-media-platforms/.

[88] Neil F Johnson et al., "Hidden Resilience And Adaptive Dynamics Of The Global Online Hate Ecology", *Nature* 573, pp. 261-265 (2019). Available at: https://www.nature.com/articles/s41586-019-1494-7.

languages."

6. **Online hate is an event-driven landscape.** Trigger events such as political elections and terrorist attacks can precipitate huge spikes in the spread of hateful narratives online.[89] Williams and Burnap analysed the emergence and propagation of online hate on Twitter in the aftermath of the Woolwich terrorist attack in 2013.[90] They found evidence of a rapid spike in online hate in the immediate aftermath of the attack which then dissipated. Other research shows similar results.[91]

7. **There are close links between online and offline forms of hate.** Online hate is often an extension of, or a precursor to, offline hate, and has the potential to amplify its harmful effects.[92] Indeed, Awan and Zempi argue that the boundaries between online and offline hate are often blurred. Some victims can find it difficult

[89] Markus Kaakinen et al., "Did The Risk Of Exposure To Online Hate Increase After The November 2015 Paris Attacks? A Group Relations Approach", *Computers In Human Behavior,* 78: 1, pp. 90-97 (2018). Available at: https://www.sciencedirect.com/science/article/abs/pii/S0747563217305484#.;
Matthew Williams et al., "Hate In The Machine: Anti-Black And Anti-Muslim Social Media Posts As Predictors Of Offline Racially And Religiously Aggravated Crime", *The British Journal Of Criminology*, 60: 1, pp. 93-117 (2020). Available at: https://academic.oup.com/bjc/article/60/1/93/5537169.

[90] Matthew Williams and Pete Burnap, "Cyberhate on Social Media in the aftermath of Woolwich: A Case Study in Computational Criminology and Big Data," *The British Journal of Criminology*, 56: 2, pp. 211-238 (2016). Available at: https://doi.org/10.1093/bjc/azv059.

[91] Bertie Vidgen, "Tweeting Islamophobia: Islamophobic hate speech amongst followers of UK political parties on Twitter – Doctoral thesis", (Oxford: University of Oxford, 2019). Available at: https://www.voxpol.eu/download/phd_thesis/Tweeting-Islamophobia-Islamophobic-hate-speech-amongst-followers-of-UK-political-parties-on-Twitter.pdf

[92] Imran Awan and Irene Zempi, "'I Will Blow Your Face Off'—Virtual And Physical World Anti-Muslim Hate Crime", *British Journal Of Criminology,* 57: 2, pp. 362-380 (2017). Available at: https://doi.org/10.1093/bjc/azv122.;
Matthew Williams et al., "Hate In The Machine: Anti-Black And Anti-Muslim Social Media Posts As Predictors Of Offline Racially And Religiously Aggravated Crime";
Karsten Müller and Carlo Schwarz, "Fanning the flames of hate; Social media and hate crime", *Journal of the European Economic Association* (2020). Available at: https://academic.oup.com/jeea/advance-article-abstract/doi/10.1093/jeea/jvaa045/5917396?redirectedFrom=fulltext.

to separate online threats from the violence and abuse they suffer offline, and live in fear of the possibility of online threats materialising in the offline world.[93] Several large-scale studies have shown that online hate is associated with offline hate crime, showing evidence of temporal and geospatial connections. However, most research uses observational datasets and further causal analyses are required to understand the mechanism by which online hate and offline attacks are linked.

8. **Online hate exhibits cross-platform dynamics.** This includes how specific bits of content, such as memes and videos, circulate across platforms, as well as how influential figures and communities migrate between platforms. For instance, several hateful communities on Reddit have created 'backup' websites for when they face quarantines or bans, including r/TheRedPill[94] and r/TheDonald.[95] Many far right influencers on mainstream platforms actively encourage their audiences to follow them on alternative platforms. The ex-leader of the British National Party, Nick Griffin, states in his Twitter biography, "Gagged here, join me on Parler and Telegram".[96]

9. **Experiences of being exposed to and targeted by online hate vary according to many factors**, including age, gender, ethnicity and religion, including age, gender, ethnicity and sexuality. It is important to acknowledge that individuals with different backgrounds and identities will have very different experiences of online hate.

[93] Imran Awan and Irene Zempi, "The Affinity Between Online And Offline Anti-Muslim Hate Crime: Dynamics And Impacts", *Aggression And Violent Behavior* 27: 1, pp. 1-18 (2016). Available at: https://www.sciencedirect.com/science/article/abs/pii/S1359178916300015.

[94] r/TheRedPill maintains a secondary website, available at: http://trp.red. Last accessed on 4 December 2020.

[95] r/TheDonald maintains a secondary website, available at: https://thedonald.win. Last accessed on 4 December 2020.

[96] Nick Griffin Twitter profile, available at: https://twitter.com/NickGriffinBU. Last accessed on 31 December 2020.

2.2.1 Online videos and hate

User-generated videos are an increasingly popular medium through which online hate is spread.[97] The length of videos means that they are well-suited to more sustained and in-depth analysis of issues and can hold viewers' attention for longer. Notably, many alt-right and alt-lite political figures from the US and the UK have been effective at attracting younger males into extremist groups through content in the same style as any viral video; high-quality editing, jump shots and leading titles.[98] Such content can attract new audiences to hateful and extremist ideas, many of whom would not traditionally identify with far-right organisations.

The ease with which hateful videos can be accessed on online platforms has raised concerns about the risk of extremist 'rabbit holes'.[99] In 2019 The New York Times reported on a self-described "brainwashed" radical, Cain Caleb, who had been indoctrinated into believing hateful ideology after viewing videos from far-right personalities on YouTube.[100] Caleb's experience, although anecdotal, demonstrates the ease with which users' beliefs can escalate when exposed to certain types of content,

[97] Gabriel Weimann and Natalie Masri, "Research Note: Spreading Hate On Tiktok", *Studies In Conflict & Terrorism*, pp. 1-14 (2020). Available at: https://www.tandfonline.com/doi/abs/10.1080/1057610X.2020.1780027.;
Matthew Barnidge et al., "Perceived exposure to and avoidance of hate speech in various communication settings", *Telematics and Informatics*, 44 (2019). Available at: https://www.sciencedirect.com/science/article/abs/pii/S0736585319307555.

[98] Angela Nagle, *Kill All Normies: Online Culture Wars from 4chan and Tumblr to Trump and the altright* (London: Zero Books, 2017).;
Bharath Ganesh, "The ungovernability of digital hate culture", *Journal of International Affairs*, 72: 2, pp. 30-49 (2018). Available at: https://www.jstor.org/stable/26552328.

[99] Raphael Ottoni et al., "Analyzing right-wing YouTube channels: hate, violence and discrimination", *Proceedings of the 10th ACM Conference on Web Science* (2018). Available at: https://arxiv.org/pdf/1804.04096.pdf.;
Maura Conway et al., "Down the (White) Rabbit hole: the extreme right and online recommender systems", *Social Science Computer Review*, 33:4, pp. 459-478 (2015). Available at: https://journals.sagepub.com/doi/pdf/10.1177/0894439314555329.

[100] Kevin Roose, "The Making of a YouTube Radical", *The New York Times*, 8 June 2019. Available at: https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html.

and is a key difference with offline settings where such self-guided radicalisation is far more difficult. The creation of extremist rabbit holes has been attributed to the role of the recommender system algorithms which present new content for users to view and engage with. Many commentators argue that recommender systems exploit individuals' initial mild preferences for, or curiosity about, certain ideologies and outlooks by showing them increasing amounts of similar (and potentially more extreme) content in an attempt to hold their attention.[101] Recommender systems are designed to optimise the content that users are shown to maximise their engagement – potentially, without any consideration for *what* users are engaging with, which increasingly has been criticised by technologists.[102] Most platforms dispute allegations that they prioritise user engagement over user safety.[103]

A further concern with videos is that their comment sections can become targets of hate. This can happen even if the video itself is not hateful. Ernst et al. analyse videos which aim to challenge Islamophobic narratives and find that they often host comments which express negative stereotypes against Muslims.[104] This is common online where any discourse can easily be hijacked and is often the case with hashtags, where activists who oppose a movement may aim to co-opt its hashtag, both to undermine the movement and

[101] Manoel Ribeiro et al., "Auditing radicalization pathways on YouTube", *Proceedings of the 2020 Conference on Fairness, Accountability and Transparency*, pp. 131-141 (2020). Available at: https://arxiv.org/pdf/1908.08313.pdf.

[102] James Williams, *Out of Our Light: Freedom and Resistance in the Attention Economy* (Cambridge: Cambridge University Press, 2018).;
Greg Elmer, "Prospecting Facebook: the limits of the economy of attention", *Media, Culture & Society*, 41: 3, pp. 332-346, (2018). Accessed at: https://doi.org/10.1177/0163443718813467;
Vikram Bhargava and Manuel Velasquez, "Ethics of the Attention Economy: The Problem of Social Media Addiction", *Business Ethics Quarterly*, pp. 1-39 (2020). Accessed at: https://doi.org/10.1017/beq.2020.32

[103] Facebook, "What 'The Social Dilemma' gets wrong", (San Francisco: Facebook, 2020). Available at: https://about.fb.com/wp-content/uploads/2020/10/What-The-Social-Dilemma-Gets-Wrong.pdf.

[104] Julian Ernst et al., "Hate beneath the counter speech?", *Journal for Deradicalization*, 10: 1, pp. 1-49 (2017). Accessed at: https://journals.sfu.ca/jd/index.php/jd/article/view/91.

to attract attention.[105] Notably, this has happened throughout the #BLM protests.[106] Mariconti et al. show that comment sections below videos can be attacked through the coordinated work of malicious actors organised on fringe websites, such as 4chan.[107] They show that troll and bot networks often operate in concert to toxify such spaces by repeatedly sharing offensive and hateful content.

2.3 Characterising and defining online hate: context and subjectivity

Defining online hate is a complex task, and a range of definitions have been put forward in previous work. Most laws focus on proscribing incitement to hatred, particularly violence. This includes the UK's 1986 Public Order Act and the EU's council Framework Decision 2008/913/JHA (both discussed in Part 1 of this report). The United Nations Commissioner for Human Rights writes that the UN "pledge[s] to publicly denounce all instances of advocacy of hatred that incites to violence, discrimination or hostility", the UN's Rabat Plan argues for "Prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence"[108] and the international human rights campaign group Article 19 advocates "Prohibiting incitement to discrimination, hostility or violence"[109]. The European Commission on Racism and Intolerance's definition of hate also focuses on incitement: "the advocacy, promotion or

[105] Kami Kosenko, "The hijacked hashtag: the constitutive features of abortion stigmatization. The #ShoutYourAbortion Twitter Campaign", *International Journal of Communication*, 13: 1, pp. 1-21 (2019). Available at: https://ijoc.org/index.php/ijoc/article/view/7849.

[106] Ryan J. Gallagher et al., "Divergent discourse between protests and counter-protests: #BlackLivesMatter and #AllLivesMatter", *Plos ONE*, 13: 4, pp. 1-23 (2018). Available at: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0195644.

[107] Enrico Mariconti et al., "You Know What to Do": Proactive Detection of YouTube Videos Targeted by Coordinated Hate Attacks".

[108] UN Office of the High Commissioner for Human Rights, *Annual report of the United Nations High Commissioner for Human Rights: addendum* (Geneva: OHCHR, 2013). Available at: https://www.ohchr.org/Documents/Issues/Opinion/SeminarRabat/Rabat_draft_outcome.pdf.

[109] Article 19, *Prohibiting incitement to discrimination, hostility or violence* (London: Article19, 2012). Available at: https://www.article19.org/data/files/medialibrary/3548/ARTICLE-19-policy-on-prohibition-to-incitement.pdf.

incitement, in any form, of the denigration, hatred or vilification of a person or group of persons, as well as any harassment, insult, negative stereotyping, stigmatisation or threat in respect of such a person or group of persons and the justification of all the preceding types of expression[.]" (p. 3)[110]

Despite the focus on "incitement to" and "stirring up" hatred in most legislative frameworks, there is a lack of consensus as to what these terms entail in practice. As Bartlett et al. argue: "[...] in respect of speech that might be deemed hateful, abusive, or racist. Defining and legislating against this type of speech is extremely difficult, and has spawned a large philosophical, linguistic, theoretical, and legal literature." (p. 11)[111] Equally, in a report for the Council of Europe on online hate moderation, published in 2020, Brown comments that "[a]n important feature of the current state of play in the governance of online hate speech is the lack of definitional harmonisation across national governments, intergovernmental organisations, Internet platforms and civil society organisations."[112] In an article with Sinclair, Brown proposes that hate is better understood as an "umbrella term" rather than a single concept.[113]

The terminological confusion apropos hate is unsurprising. It is what the philosopher W. B. Gallie describes as an "essentially contested concept" – numerous definitions and accounts proliferate, with relatively little consensus about the core features.[114] Part of the challenge is that definitions are, by their nature, stipulated at a high level. Further analysis and theoretical discussion is needed to clarify what terms such as "incitement to" and

[110] European Commission against Racism and Intolerance (ECRI), *ECRI General Policy Recommendation No. 15 on Combating Hate Speech* (Strasbourg: Council of Europe, 2016). Available at; https://rm.coe.int/ecri-general-policy-recommendation-no-15-on-combating-hate-speech/16808b5b01.

[111] Jamie Bartlett et al., *Anti-social media* (London: Demos, 2014). Available at: https://www.demos.co.uk/files/DEMOS_Anti-social_Media.pdf.

[112] Alexander Brown, *Models of Governance of Online Hate Speech* (Brussels: Council of Europe, 2020). Available at: https://rm.coe.int/models-of-governance-of-online-hate-speech/16809e671d.

[113] Alexander Brown and Adriana Sinclair, *The Politics of Hate Speech Laws* (London: Routledge, 2019).

[114] W.B Gallie, "Essentially Contested Concepts," *Proceedings of the Aristotelian Society* 56:1, pp. 167-98 (1955). Available at: https://www.jstor.org/stable/4544562.

"stirring up" actually mean. For example, whether content is considered hateful depends in large part on what words and/or imagery it contains. It also depends on the *context* in which it is shared. As Parekh puts it, "[e]very aspect of online hate – its moral and emotional significance, content, import and insinuations – are inseparable from, and can only be determined in light of, context." (p.42)[115] From a more practical perspective, this sentiment is echoed in the CPS guidance on prosecuting online hate: "Each case must be decided on its own facts and merits and with particular regard to the context of the message concerned."[116] Context can operate in many different ways and several frameworks for explicating its role have been put forward (See Box 2). These primarily focus on the role of the speaker, the audience, the broader social and historical context and the form of the content, such as whether its modality and medium.

Box 2: Frameworks for understanding the role of context in hate		
The UN's Rabat Plan of Action identifies six elements which all need to be met for a hateful statement to amount to a criminal offence. It accounts for: 1. Context of the statement	The CPS's guidance on online hate outlines several procedural aspects of context: ● Who is the intended recipient? ● Does the message refer to their characteristics?	Harvard academic Benesch proposes five elements which can make language "dangerous" and emphasises five aspects of context in her work: 1) The speech act itself 2) The audience

[115] Bhikhu Parekh, "Is there a case for banning hate speech?", pp. 37–56 in M. Herz and P. Molnar (eds.) *The content and context of hate speech: Rethinking regulation and responses* (Cambridge: Cambridge University Press, 2012).

[116] Crown Prosecution Service, *Social Media - Guidelines on prosecuting cases involving communications sent via social media* (London: Crown Prosecution Service, 2018). Available at: https://www.cps.gov.uk/legal-guidance/social-media-guidelines-prosecuting-cases-involving-communications-sent-social-media.

2. Speaker's position or status 3. Intent to incite the audience against a target group 4. Content and form of the speech 5. Extent of its dissemination 6. Likelihood of harm, including imminence[117]	● Can the nature of the message be understood with reference to a news or historical event? ● Are terms which require interpretation, or explanation by the recipient, used? ● Was there other concurrent messaging in similar terms so that the suspect knowingly contributed to a barrage of such messages?[118]	3) Social and historical context 4) The speaker 5) Mode of dissemination[119]

Context is a complex issue and can be understood in many different ways. For clarity, we focus on two main aspects. First, context affects whether content should be considered

[117] UN Office of the High Commissioner for Human Rights, *Rabat Threshold Test* (New York: OHCHR, 2020). Available at: https://www.ohchr.org/Documents/Issues/Opinion/Articles19-20/ThresholdTestTranslations/Rabat_threshold_test.pdf.
[118] Crown Prosecution Service, *Social Media - Guidelines on prosecuting cases involving communications sent via social media*.
[119] Susan Benesch, *What is Dangerous Speech?* (Washington: Dangerous Speech, 2020). Available at: https://dangerousspeech.org/about-dangerous-speech/.

hateful. This is primarily related to the identity of the content creator,[120] the importance of which is demonstrated in the polysemic use of 'pejorative' terms and slurs. These are often reclaimed by individuals who have either personally been targeted by hate or who are from groups and communities which have been targeted.[121] For example, the racist term 'n*gga' has been widely reappropriated by black people, the homophobic term 'qu*er' by gay communities, and misogynistic terms 'c*nt' and 'b*tch' by women. Such language has a fundamentally different meaning when it is reclaimed and used by the targeted groups: it is no longer hateful as the reappropriation "undermin[es] the signal strength of the slurring term".[122]

Note that this is a very sensitive issue and whether the use of a slur either "reinforces" or "subverts" status hierarchies depends on a range of factors.[123] For instance, some 'allies' of groups who are targeted by hate may feel comfortable using reclaimed hateful slurs but this can be challenged by others who either are unaware that they are allies or do not feel comfortable with their use. Further, not all slurs have been reclaimed and some language remains overwhelmingly hateful in its use, such as terms like "r*tard", which target people with disabilities. As Bolinger puts it, even though all slurs *can* be reappropriated, "that's not to say that a reclaimed slurring-term can be used by just anyone, or in just any context, without warranting offence."[124]

Second, context affects the impact of online hate and the harm that it causes. If the person who creates and/or shares hate is a powerful figure then they will have more

[120] Conor O'Dea and Donald Saucier, "Perceptions of racial slurs used by black individuals towards white individuals: derogation or affiliation?", *Journal of Language and Social Psychology*, 39:5-6, pp. 678-700 (2020). Available at: https://doi.org/10.1177/0261927X20904983.

[121] Bianca Cepollaro and Dan Zeman, "The challenge from non-derogatory uses of slurs", *BRILL*, 97: 1, pp.1-10 (2020). Available at: https://doi.org/10.1163/18756735-09701002.

[122] Renée Jorgensen Bolinger, "The Pragmatics of Slurs", *Nous*, 51: 3, pp. 439-462 (2017). Available at: https://doi.org/10.1111/nous.12090.

[123] Conor O'Dea and Donald Saucier, "Perceptions of racial slurs used by black individuals towards white individuals: derogation or affiliation?".

[124] Renée Jorgensen Bolinger, "The Pragmatics of Slurs".

rhetorical power and so their content is likely to have far greater impact. Equally, some settings are highly conducive to spreading hateful content. In periods of heightened tensions, such as in the aftermath of a terrorist attack, it is likely harm will be amplified.[125] This is discussed below in relation to content's influence.

As well as being contextual, hate can also be perceived in different ways and as such is best understood as *subjective*. Multiple studies have shown that interpretations of hatefulness differ substantially across individuals, particularly with content which is more ambiguous.[126] There is a lack of systematic large-scale research into *why* these differences exist and what factors affect perceptions of hate. The limited available evidence indicates that traits such as political affiliation and age play a key role[127] as well as an individuals' gender, political outlook and personality traits.[128] Research conducted on British millennials' perceptions of hate speech showed that they typically adopt a bolder, more radical definition than older generations. The interviewees defined themselves as generation 'woke' and the researchers summarised that from younger generation's perspective, racist hate speech, "[...] no longer regarded as comprising racial slurs alone, but also as including post-colonial nuances. Furthermore, interviewees incorporated stigma against sexuality, gender and particularly transgender rights in their

[125] Matthew Williams, *Hatred Behind the Scenes* (London: Mishcon de Reya, 2020). Available at: https://hatelab.net/wp-content/uploads/2019/11/Hatred-Behind-the-Screens.pdf.

[126] Joni Salminen et al., "Online hate ratings vary by extremes: a statistical analysis", *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, pp. 213-217 (2019). Available at: https://doi.org/10.1145/3295750.3298954.

[127] Matthew Costello et al., "Social Group Identity And Perceptions Of Online Hate", *Sociological Inquiry*, 89: 3, pp. 427-452 (2019). Available at: https://doi.org/10.1111/soin.12274.

[128] Alison Bacon et al., "Understanding public attitudes to hate: developing and testing a U.K. version of the hate crime beliefs scale", *Journal of Interpersonal Violence*, 0:0, pp. 1-26 (2020) Available at: https://doi.org/10.1177/0886260520906188;
Daniel Downs et al., "Predicting the Importance of Freedom of Speech and the Perceived Harm of Hate Speech", *Journal of Applied Social Psychology*, 42:6, pp. 1353–1375 (2012) Available at: https://doi.org/10.1111/j.1559-1816.2012.00902.x.

definition of hate speech." (p.58)[129] However, the explanatory power of such studies is low, with one paper noting that interpretation of whether content is hateful "differs more by individual than by the country [they are from]".[130]

The contextual and subjective aspects of hate are often conflated – but it is important to separate them given that they capture different aspects of hate, and different reasons for why definitions of hate are often contested. Context captures intra-individual variation and how any one person may view the same content differently depending on who produces it, when and to whom. Subjectivity captures inter-individual variation and how different people can view the same exact content in the same exact context differently.

2.4 How online hate inflicts harm

Online hate speech can inflict multiple harms, many of which are similar to the harms created by other forms of undesirable and restricted online content. For instance, in a 2020 report the civil society organisation Hope Not Hate describes both emotional and physical harms "that occur on, or are facilitated by, the online world" (p. 6).[131] The Law Commission's 2020 scoping report on hate crime identifies the "emotional and psychological harms of hate speech" as well as the "social exclusion and marginalisation of vulnerable groups in society" which it can lead to.[132] A large body of academic work has also explored the issue.

[129] Stavros Assimakopoulos et al., "Young People's Perception of Hate Speech", pp. 53-85 in Stavros Assimaopoulos et al. (eds.), Online Hate Speech in the European Union: *A Discourse-Analytic Perspective* (Berlin: Springer, 2017). Available at: https://link.springer.com/chapter/10.1007/978-3-319-72604-5_4.
[130] Joni Salminen et al., "Online Hate Interpretation Varies by Country, But More by Individual: A Statistical Analysis Using Crowdsourced Ratings," in *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)* (Valencia: IEEE), pp. 88-94 (2018). Available at: https://ieeexplore.ieee.org/document/8554954.
[131] HOPE not Hate, *A Better Web: Regulating to Reduce Far-Right Hate Online* (London: HOPE not Hate, 2020). Available at: https://www.hopenothate.org.uk/wp-content/uploads/2020/11/A-Better-Web-Final-.pdf.
[132] The Law Commission, *Hate Crime: Consultation Paper Summary.*

Matsuda et al.'s 1993 edited book Words that Wound: Critical Race Theory, Assaultive Speech and the First Amendment is a landmark text in hate studies for drawing widespread attention to the harm of hateful language.[133] They describe the metaphorical "gut punch" that hate can inflict (p. 23), arguing that it creates unjustifiable and life-impacting feelings of anxiety and fear amongst victims – and that its affective power comes from practices of oppression and discrimination that operate in society, rather than individuals' resilience (or perceived lack thereof). Queer theorist Butler has further explored how hate can inflict harm, which she argues is a result of the broader social and linguistic structures in which communications are embedded. For instance, she argues that every use of a racist slur by a white person invokes a history and legacy of oppression, giving them resonance and force which transcends the experiences of any single individual.[134] Later scholars such as Benesch have extended these theories on how the power of language relates to the wider social context.[135]

Debates about the harm in hate speech were reinvigorated in 2012 with the publication of Waldron's aptly titled The Harm in Hate Speech.[136] His work has attracted attention for arguing that minimising the degree to which individuals feel offended is not sufficient justification to regulate hate speech. Instead, he argues that online hate should be regulated because it undermines the dignity of groups, impinging on their rights and their ability to engage democratically, feel safe and to take part in civic discourse. This position has attracted considerable support, although many have debated whether harm is a

[133] Mari Matsuda et al. (eds.) *Words that Wound: Critical Race Theory, Assaultive Speech and the First Amendment* (New York: Westview Press, 1993).
See also: Timothy Jay, "Do offensive words harm people?" *Psychology, Public Policy, and Law,* 15: 2, pp. 81–101 (2009). Available at: https://doi.org/10.1037/a0015646.

[134] Judith Butler, *Excitable Speech: A Politics of the Performative* (London: Routledge, 1997).

[135] Susan Benesch, *What is Dangerous Speech?*

[136] Jeremy Waldron, *The Harm in Hate Speech* (Oxford: Oxford University Press, 2012).

consequence of hate or a *constitutive* feature, which remains somewhat unclear in Waldron's work.[137]

Drawing on relevant empirical and theoretical work we outline seven ways in which online hate can inflict harm (See Box 3). All hate has the potential to inflict all of these harms – but which harms manifest will depend heavily on both the content and the context in which it appears. The harms inflicted by hate are primarily borne by individuals who are directly targeted and the wider communities which they are from. However, it is worth noting that 'bystanders' and 'allies' can also experience harm from observing hate online, although this is generally to a far smaller degree and the primary focus is the direct targets.

Box 3: The harm caused by online hate[138]	
1	The **immediate distress and emotional harm** that individuals can experience when viewing, or being targeted by, hateful content. The harm may be heightened if the individual has been targeted by online hate previously.
2	The **long-term mental health effects** of being targeted by online hate, particularly if this is combined with other forms of harmful behaviour, such as stalking and/or harassment.

[137] Eric Barendt, "What is the Harm of Hate Speech?", *Ethical Theory and Moral Practice*, 22:3, pp. 539-553 (2019). Available at: https://doi.org/10.1007/s10677-019-10002-0.

[138] The authors thank Josh Cowls for his helpful discussions and advice on the harm inflicted by online hate.

3	The **long-term impact on victims' behaviour**. Being targeted by online hate can lead individuals to change how they live their lives. In some cases, individuals report not wanting to leave the house out of fear.[139]
4	The **negative effect on individuals' willingness to engage in public and civic forums** and discussions, such as taking on prominent public positions. This is possibly one of the most pernicious effects of online hate, and inflicts harm at three levels: 1. The individuals who are unwilling to enter public and civic life may suffer personally. 2. The groups to which they belong will suffer if their perspective is not articulated publicly. 3. Society as a whole misses out on a plurality of perspectives. The basis of democracy is robust debate; if certain groups are excluded then everyone loses out from having less relevant, engaged, diverse and critical discussions.
5	**Motivating and enabling offline attacks and other forms of harm**. Some hateful content directly calls on its audience to attack minority groups, whereas in other content such calls are implicit or not present and the content is best understood as 'inspiring' rather than 'inciting' harm (see above) – nonetheless, the ideas and views expressed in hateful online content may still lead parts of the audience to inflict harm on victims in an offline setting. Whether online hate leads to offline attacks depends heavily on the setting, and further research is still needed.[140]

[139] Runnymede Trust, *Islamophobia: Still a challenge for us all*, (London: The Runnymede Trust, 2017).

[140] See: Karsten Müller and Carlo Schwarz, "Fanning the flames of hate; Social media and hate crime"; Matthew Williams et al., "Hate In The Machine: Anti-Black And Anti-Muslim Social Media Posts As Predictors Of Offline Racially And Religiously Aggravated Crime".

6	**Motivating and enabling other online attacks**. Exposure to online hate can inflict other online harms: a person who views hateful content may become motivated to target individual members of a group. This could include financial attacks or scams, hacking them, doxing them (i.e., where individuals are attacked by having their private and personally identifying information shared online), or using a so-called ‘honeypot’ to motivate the victim to engage in criminal activity (for which they could subsequently be prosecuted). Due to the sensitive nature of these other online attacks, and their resource intensiveness, this remains an under-researched area.
7	The **implications for social justice and fairness** of tolerating online hate against some groups. This is the least tangible form of hazard but is important: a society in which already-marginalised and vulnerable groups are routinely harassed and attacked raises fundamental questions about its fairness.

2.5 Substance: what online hate expresses

Online hate is highly varied. To provide clarity, we split hateful content into four types. This typology reflects prior social, critical and computational research in hate studies.[141] Other typologies have been put forward which make similar albeit subtly different distinctions. Subtypes with descriptions are given in Box 4. Note that these types are not

[141] Bertie Vidgen et al., “Challenges and frontiers in abusive content detection” in *Proceedings of the Third Workshop on Abusive Language Online* (Florence: Association for Computational Linguistics), pp. 80-93 (2020). Available at: https://www.aclweb.org/anthology/W19-3509/.;
Iginio Gagliardone et al., *Countering online hate speech.*
Manfred Kienpointner, “Impoliteness online, hate speech in online interactions”, *Internet Pragmatics*, 1: 2, pp. 329-351 (2018). Available at: https://www.jbe-platform.com/content/journals/10.1075/ip.00015.kie.;
Imran Awan, “Islamophobia and Twitter: a typology of online hate against Muslims on social media”, *Policy & Internet*, 6: 2, pp. 133-150 (2014). Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/1944-2866.POI364.

based on legal definitions. In practice, content may not neatly fit into one type or subtype, or could involve several different types and subtypes at once.

1. **Threatening:** content which expresses intention to engage in harmful actions against a group or members of a group on the basis of their group identity.[142] It includes threats of physical violence.

2. **Inciting**: content which explicitly encourages, advocates or justifies harm to be inflicted on a group or members of a group on the basis of their group identity. It includes incitement to physical violence.

3. **Demonising:** content which is explicitly hateful but does not involve threats or incitement. It is likely to *inspire* hatred in others and thus may have similar harmful effects. It includes derogating, demeaning, insulting and attacking a group.

4. **Animosity:** content which *expresses* prejudice against a group but does not explicitly attack them. It includes content which others a group[143] by emphasising their difference, strangeness or unimportance or by mocking and undermining their experiences. It is usually less 'strong' and is more likely to be expressed without intention to inflict harm.

Box 4: Types and subtypes of hateful content
Threatening • Expressing intention to personally engage in action/violence against members of an identity. This includes a user directly threatening another user on the basis of their identity or saying that they intend on harming members of a groups.

[142] Note that this wording is used to highlight an important point: threats against an individual are not hateful if the attack does not reference or is motivated by the individuals' group membership.

[143] Gordon Allport, *The nature of prejudice* (New York: Addison-Wesley, 1954).

Inciting

- Inviting/encouraging action/violence to be taken against members of an identity. This includes directly requesting that others plan and organise an attack or asking for their assistance in planning/conducting harmful activities.
- Presenting normative justification for action/violence to be taken against members of an identity. This includes making a moral case for inflicting harm upon certain groups, such as claiming they deserve to be attacked.
- Defending/legitimising action/violence which has been taken against members of a group. This can include celebrating past terror attacks, offline hate crimes or other hateful events.

Demonising

- Dehumanising statements about members of a group. This includes describing, or comparing, a group as trash, subhuman, waste or a disease.
- Explicit attacks which demean an identity. This includes being insulting to, or malicious about, an identity, such as describing them as weak, degenerate or unwelcome.
- Negatively comparing one group with another, typically a socially dominant in-group, and/or the rest of society.
- Stating that a group is not welcome. This can be particularly hateful if it undermines the legitimate claims to citizenship of a group who, although native to a country, may be portrayed as foreigners due to their religion or ethnicity.
- Portraying a group as a threat.
- Use of slurs and some pejoratives.

Animosity

- Emphasising the unusualness and difference of another group, constructing them as an alien and out-of-place 'Other'. This includes stating that groups are strange or mocking their cultural practices and habits.
- Hateful jokes and satire, such as lampooning the practices, speech or beliefs of a group. Whether jokes made about groups are actually hateful has been extensively debated. One useful consideration is whether the joke is made to

undermine or uplift the group in question; jokes which undermine the group are likely to fall into animosity.[144]

- Defence of prejudice. This includes morally justifying hate through defence of freedom of speech or proposing other societal benefits of prejudicial content. Whether defence of prejudice is itself hateful will depend heavily upon the nature of the defence that is given.
- Interactional forms of prejudice, such as only responding to comments from people who are of the same group. This is often called micro-aggression and is usually not addressed through legal or regulatory frameworks.[145]
- Use of some pejoratives may qualify as animosity rather than demonisation.

Not hateful

- Critical and neutral discussions of groups.
- Counterspeech.
- Attacks against abstract concepts and institutions.
- Uncivil and profane language.

Drawing the line between different types of hateful content can involve making contentious and sometimes-subjective choices. To provide further clarity we note the following decision boundaries:

1. Threatening and inciting content can be understood in relation to *action*. In both cases, the content relates to a harm that could be inflicted on the targeted group or its members. Whether or not the harm will actually be inflicted is of secondary importance; both types raise the possibility of harm being inflicted and would be

[144] Simon Weaver, “A rhetorical discourse analysis of online anti-Muslim and anti-Semitic jokes”, *Ethnic and Racial Studies*, 36: 3, pp. 483-499 (2011). Available at: https://www.tandfonline.com/doi/abs/10.1080/01419870.2013.734386.

[145] Rob Eschmann, “Digital Resistance: How online communication facilitates responses to racial microaggressions”, *Sociology of Race and Ethnicity*, pp. 1-14 (2020). Available at: https://doi.org/10.1177/2332649220933307.

likely to make victims fear such an outcome. In contrast, demonising and animosity do not directly involve action.

2. Separating inciting from demonising content can be difficult as it depends on how far one interprets the *implied* action that may be expressed in demonising content. The lynchpin of the difference is whether the content calls for harm to be inflicted on a group. If such a call is made then it is likely that the content *incites* harm. If the content does not explicitly encourage, advocate or justify harmful actions (nor expresses intention to engage in harmful action) then it is likely that it *inspires* hate. Such content may still be harmful in other ways, but the key point is that it inflicts such harm through making hateful statements about a group – and not through directly inciting harm against them.

3. Animosity is the most likely to be edge case content. This is content which falls on or near to the boundary between hate and non-hate. It is typically highly contested and people are likely to have very different perspectives on whether it is harmful and how it should be handled. Some may see animosity as clearly hateful whereas others will see it as what Imhoff and Recker call "legitimate critique" (in relation to Islamophobia): content that is critical or questioning but not hateful.[146] Removing content in the animosity type raises the greatest risk that individuals' freedom of expression will be constrained, especially given possible mistakes in moderation. Box 5 shows an example of how an edge case involving the conservative commentator Steve Crowder was handled by YouTube.

[146] Roland Imhoff and Julia Recker, "Differentiating Islamophobia: Introducing A New Scale To Measure Islamoprejudice And Secular Islam Critique", *Political Psychology*, 33: 6, pp. 811-824 (2012). Available at: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1467-9221.2012.00911.x.

Box 5: Steve Crowder on YouTube

In 2019 YouTube stopped conservative commentator and vlogger Steven Crowder from running ads on his channel. This was due to reports from Vox writer and video host Carlos Maza that Crowder had directed homophobic language against him.[147] Crowder had also sold t-shirts on his website which featured a homophobic slur. Initially, the platform responded to the allegations by claiming that Crowder did not violate any of its policies:[148]

> "Our teams spent the last few days conducting an in-depth review of the videos flagged to us, and while we found language that was clearly hurtful, the videos as posted don't violate our policies. We've included more info below to explain this decision:
>
> As an open platform, it's crucial for us to allow everyone–from creators to journalists to late-night TV hosts–to express their opinions w/in the scope of our policies. Opinions can be deeply offensive, but if they don't violate our policies, they'll remain on our site.
>
> Even if a video remains on our site, it doesn't mean we endorse/support that viewpoint."[149]

[147] Julia Alexander, "YouTube revokes ads from Steven Crowder until he stops linking to his homophobic T-shirts", *The Verge*, 5 June 2019. Available at: https://www.theverge.com/2019/6/5/18654196/steven-crowder-demonetized-carlos-maza-youtube-homophobic-language-ads.

[148] Nick Statt, "YouTube decides that homophobic harassment does not violate its policies", *The Verge*, 4 June 2019. Available at: https://www.theverge.com/2019/6/4/18653088/youtube-steven-crowder-carlos-maza-harassment-bullying-enforcement-verdict.

[149] TeamYouTube (@TeamYouTube) Twitter post, "Our teams spent the last few days conducting an in-depth review of the videos flagged to us, and while we found language that was clearly hurtful, the videos as posted don't violate our policies" (San Francisco: Twitter, 5 June 2019). Available at: https://twitter.com/TeamYouTube/status/1136055351885815808. Last accessed on 4 December 2020.

After facing increasing pressure from the public, YouTube demonetised Crowder's channel, stating, "We came to this decision because a pattern of egregious actions has harmed the broader community and is against our YouTube Partner Program policies."[150] However, YouTube did not remove Crowder's channel completely and reported that the demonetisation was not permanent: privileges would be restored if he could "address all of the issues", including removing links to his store that sells homophobic t-shirts.[151] After over a year of suspension, Crowder was reinstated in the Partner Program in August 2020 and thus allowed to run ads on his channel again.[152]

During this time, YouTube updated its harassment policy to take a stronger stance on personal attacks and prejudice-driven ad hominem insults. In December 2019 it announced, "We will no longer allow content that maliciously insults someone based on protected attributes such as their race, gender expression, or sexual orientation. This applies to everyone, from private individuals, to YouTube creators, to public officials."[153]

[150] TeamYouTube (@TeamYouTube) Twitter post, "We came to this decision because a pattern of egregious actions has harmed the broader community and is against our YouTube Partner Program policies" (San Francisco: Twitter, 5 June 2019). Available at: https://twitter.com/teamyoutube/status/1136341801109843968?lang=en. Last accessed on 4 December 2020.

[151] TeamYouTube (@TeamYouTube) Twitter post, "Sorry for the confusion, we were responding to your tweets about the T-shirt" (San Francisco: Twitter, 5 June 2019). Available at: https://twitter.com/teamyoutube/status/1136363701882064896?s=21. Last accessed on 4 December 2020.

[152] Julia Alexander, "Youtube Will Let Steven Crowder Run Ads After Year-Long Suspension For Harassment", *The Verge*, 12 August 2020. Available at: https://www.theverge.com/2020/8/12/21365601/youtube-steven-crowder-monetization-reinstated-harassment-carlos-maza.

[153] Matt Halprin, "An update to our harassment policy, Official YouTube Blog", *YouTube*, 11 December 2019. Available at: https://blog.youtube/news-and-events/an-update-to-our-harassment-policy/.

2.6 Articulation: how online hate is expressed

As well as varying in terms of its substance, hateful online content can vary in how it is articulated. Some hate will involve amplifying elements, such as swear words, which heighten the aggression and vitriol, and make the hate far more obvious. Other content will be harder to identify as it might be more covert or intentionally expressed with ambiguity. In some cases, malicious actors may aim to intentionally obfuscate their content, thereby making it harder for content moderation systems to detect. More covert forms of hateful language often rely more heavily on context (see Section 2.3). For instance, 'dog whistles' are seemingly innocuous statements that communicate a hateful message to a select portion of the audience (i.e., the 'dogs' who can hear the whistle). They have been widely used by far right politicians to circumvent any restrictions on what can be said, particularly when making public statements.[154] Individuals unfamiliar with hateful discourses may miss such remarks entirely and view them as legitimate forms of critical discussion. Box 6 shows different ways in which online hate is articulated, ranging from the most covert to the most overt.

Box 6: The articulation of online hate, from covert to overt[155]			
Harder for humans to identify	**Harder for automated software to identify**	**Straightforward to identify**	**Harder to miss**

[154] Prashanth Bhat and Ofra Klein, "Covert Hate Speech: White Nationalists and Dog Whistle Communication on Twitter", pp.151-172 in Gwen Bouvier and Judith Rosenbaum (eds.) *Twitter, The Public Sphere and the Chaos of Online Deliberation* (Switzerland: Palgrave MacMillan, 2020).

[155] All examples of online hate in the report are synthetic. We aimed to ensure that they are broadly realistic given content we have observed online. We also aimed to minimise ideological expressions of hate, although in some cases we felt this was necessary to ensure realism.

Convoluted statements and complex forms of language; dog whistles.	**Intentional mis-spellings, such as elongations, use of punctuation, letter replacements and homophonic spellings.**	**Explicit remarks and overt statements, such as use of slurs.**	**Use of ‘intensifying’ language and aggressive content.**
“I just wish that fewer of them lived round here, they make everything feel a lot darker :p” “Butterflies aren’t welcome. I’m like a catcher, Ima crush any twinkly mf that I see.” “(((they))) control the media, just wait until they get their own back for what AH tried to complete. Be on your guard, don’t trust them. 88”	“I dnt l!ke blayuk ppl on ma streeeeet” “G*yz are not not not allowed round here, although sometimes we let them in so we can give em a beeeeeting. “You cant trust a joooooooo big nose, they are scom.”	”I don’t like black people” “Gay people make me feel uncomfortable, I don’t like being near them.” “Jewish people are untrustworthy”	“I FUCKING hate niggas” “SCUM. Sick of these pedo nonce gays down my street, they got married, what will be next? Incest??!” “Bloody jews, I hate them. Just stick them in an oven, they’re shit bag animals”

The substance and articulation of online hate often overlap and intersect; content which uses swear words and vitriol is not only ‘more obvious’, it is also often more harmful and likely to evoke a greater emotional response from victims. Nonetheless, in most cases it is useful to separate *what is said* (substance) from *how it is said* (articulation).[156] This distinction also has implications for how online hate is moderated. Platforms may decide to develop different technologies and processes to address the different challenges posed by (a) content which is more overt and as such easier to identify and take down (by both humans and AI) and also (b) content which expresses deeply hateful views but is articulated in a more nuanced or obfuscated way.

2.7 The hazard and influence of online hate

To better understand how any bit of hateful content harms (see Section 2.4), we analyse it in terms of **hazard**, which we define as:

> The *potential* of content to inflict harm in a given context.

Hazard is a combination of the hateful content's substance (i.e., what is expresses) and its articulation (how it is expressed). Evaluating hazard requires understanding the intrinsic features of the content, such as its type (see Section 2.5) and whether it is overt or covert (see Section 2.6), in conjunction with the broader social, historical and political context. The degree of harm inflicted by hateful online content depends not only the content's hazard but also on how many individuals are exposed to and/or targeted by it, and the power that it has over them. This is linked to many of the contextual aspects of online hate discussed above, including the authority of the speaker, the broader social backdrop and the receptibility of the audience, as well as the socio-technical affordances of how the content is shared, such as the platform design and the medium. Accordingly, to understand its potential harm, content's hazard must be assessed in conjunction with an additional factor: its **influence**, which we define as:

> The reach and resonance of hateful content.

[156] Patrícia Rossini, "Beyond Incivility: Understanding Patterns of Uncivil and Intolerant Discourse in Online Political Talk," *Communication Research* (2020). Available at: https://doi.org/10.1177/0093650220921314.

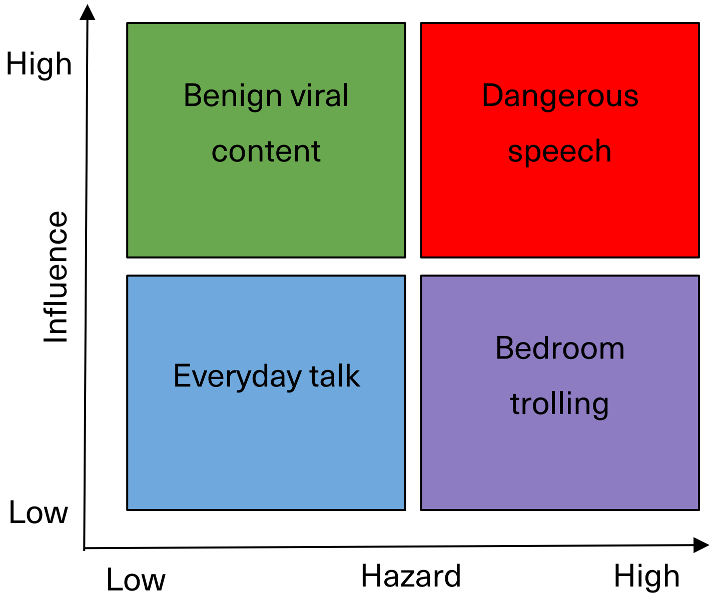

Figure 1: The likely harm of online hate is a product of hazard and influence.

Figure 1 shows how hazard and influence can be used to understand the potential harm inflicted by hate. The x-axis (along the bottom) shows the degree of hazard, ranging from low (e.g., content which is neutral) to high (e.g., threatening content). The y-axis (along the vertical side of Figure 1) shows the degree of influence that the content has, ranging from low to high. We have split the different configurations of online hate's hazard and influence into four quadrants, showing four archetypes of content. The four quadrants are:

1. **Dangerous speech** is highly hazardous content which has substantial influence. It is seen by and negatively impacts many people.

2. **Bedroom trolling** is content which is highly hazardous, as with dangerous speech, but in contrast reaches and influences very few people. It can still cause substantial harm to those individuals, but there are far fewer of them. Note that we call this content 'bedroom' trolling to capture the limited nature of its reach, rather than to indicate where it is created.

3. **Benign viral content** comprises neutral and positive messages which are seen by many people. They may be entirely unrelated to issues of identity and hate.

4. **Everyday talk** is content which does not contain anything hateful and has very little influence because very few people see or engage with it. Most content will fall into this category; a large body of research shows that most online content is either not engaged with at all or by only a small number of people online.[157]

In most settings, particularly with content moderation, online hate is evaluated *a priori* on the basis of its likely hazard and influence (and as such its likely harm). This creates what we term the **harm paradox***:*

> Most content is assessed based on what it expresses and how it is expressed rather than its empirical effects. As such, it is often unknown whether content that is labelled 'harmful' has actually inflicted harm.

For instance, in certain settings covert animosity could have a severely negative impact on victims, whereas overt threatening language may inflict little actual harm if it is viewed by very few people. In this sense, the harm inflicted by online hate is not deterministic but inherently somewhat random. This is due in part to the uncertainty of online settings, which make it very hard to predict the impact and reach of content.[158] In contrast with offline environments where content diffusion depends on hard-to-access infrastructure controlled by gatekeepers, such as being invited on radio and TV shows, in a networked environment any speakers' content can potentially go viral and be witnessed by an

[157] Sharad Goel et al., "The Structural Virality of Online Diffusion", *Management Science*, 62: 1 (2015). Available at: https://doi.org/10.1287/mnsc.2015.2158

[158] Helen Margetts et al., *Political Turbulence: How Social Media Shape Collective Action*, (Oxford: Princeton University Press, 2016).

audience far larger and more varied than intended or expected.[159] Put simply, the **harm paradox** captures the fact that in most cases there is likely to be a gap between (a) how the potential of content to inflict harm has been assessed (i.e., its hazard combined with its influence) and (b) the actual harm that it would inflict. The size of this gap will vary from case-to-case and by nature is difficult to evaluate.

[159] Alice Marwick and danah boyd, "I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience", *New Media & Society*, 13: 1, pp. 114-133 (2012). Available at: https://doi.org/10.1177/1461444810365313;
Geah Pressgrove et al., "What is Contagious? Exploring why content goes viral on Twitter: A case study of the ALS Ice Bucket Challenge", *International Journal of Nonprofit and Voluntary Sector*, pp. 1-8 (2020). Available at: https://doi.org/10.1002/nvsm.1586

Part 3: Addressing online hate

3.1 Content moderation for online hate

Online hate is primarily addressed by VSPs, as well as other online platforms, by content moderation systems. These can be understood as "[t]he organised practice of screening user-generated content posted to Internet sites, social media and other online outlets, in order to determine the appropriateness of the content for a given site, locality, or jurisdiction." (p. 1)[160] In recent years, the design and governance of content moderation systems has emerged as a key focal point in debates about users' safety online; in a 2020 review paper Gillespie et al. argue that "Content moderation [...] has exploded as a public, advocacy, and policy concern".[161] Content moderation systems are best understood as socio-technical systems, in which the policies and values of the platform shape the technology and processes that are implemented – but the affordances of the technology also shape what is considered possible and how policies are implemented.[162] Large VSPs have a substantial infrastructure of people, process and technology to both shape and implement design choices.[163] Creating an effective moderation system is likely to require teams with expertise from a range of subject matters and domains, including ethics, engineering, social science, machine learning and policy.

The failure points of technical systems often reflect the values and biases of their creators. For instance, in a landmark study Buolamwini and Gebru examined the gender-

[160] Sarah T. Roberts, "Content moderation", in Larry Schintler and Clea McNeel (eds.), *Encyclopaedia of Big Data* (Berlin: Springer, 2017).

[161] Tarleton Gillespie et al., "Expanding the Debate about Content Moderation: Scholarly Research Agendas for the Coming Policy Debates." *Internet Policy Review*, 9: 4 (2020). Available at: https://doi.org/10.14763/2020.4.1512

[162] Dubravka Cecez-Kecmanovic, "The sociomateriality of information systems: current status, future directions", *MIS Quarterly,* 38: 3, pp. 809-830. Available at: https://doi.org/10.25300/MISQ/2014/38:3.3

[163] Tarleton Gillespie, "Content moderation, AI and the question of scale", *Big Data & Society*, July-December, pp. 1-5, (2020). Available at: https://doi.org/10.1177/2053951720943234

and race- biases of facial recognition algorithms.[164] They found that the datasets used to train facial recognition systems lacked darker-skinned females and women, leading to far higher error rates on these groups, especially at the intersection (i.e., darker-skinned women suffered from particularly high error rates).

What makes a 'good' content moderation system has been heavily debated, with a wide array of design principles being put forward in academic research. In practice, the requirements of specific moderation system are likely to vary based on the users and uses of a platform – and it is crucial that a 'one size fits all' approach is avoided when considering how to design or improve content moderation. Nonetheless, some recurrent themes appear across previous literature. The requirements of an effective content moderation system also overlap with the features of other socio-technical systems. For instance, a report on The Ethics of AI from the Alan Turing Institute argues that systems need to be fair, accountable, sustainable and transparent; all principles which are relevant for content moderation systems.[165]

We identify five desirable features of effective content moderation systems. They are agnostic to whether the system is primarily human-driven (i.e., through manual content review) or technology-driven (i.e., through the use of AI).

- **High performing**: Systems that can correctly identify hateful and non-hateful content. This can be evaluated through the system's precision (i.e., how much of the content identified as hateful actually is hateful) and recall (i.e., how much of all the hateful content has been identified). One concern is that even high-performing systems may become fast-outdated and this must be evaluated on an ongoing basis.

[164] Joy Buolamwini and Timnit Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification", *Proceedings of Machine Learning Research*, 81, pp. 1-15 (2018). Available at: http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

[165] David Leslie, *Understanding artificial intelligence, ethics and safety* (London: The Alan Turing Institute, 2019). Available at: https://www.turing.ac.uk/research/publications/understanding-artificial-intelligence-ethics-and-safety.

- **Fair**: Systems that work equally well across different groups. How fairness should be defined and evaluated is increasingly debated.[166] Typically, performance metrics (e.g. recall, precision) are compared across different groups. A fair system is one that has an equal rate of error or rate of identification. Notably, Davidson et al. and Saps et al. show that AI-based hate and toxicity detection systems have different levels of accuracy on content produced by different social groups, with a far higher rate of error on content produced in African-American vernacular.[167]

- **Robust**: Systems that are capable of withstanding adversarial attacks, are robust to minor variations in content and work equally well on different types and media of content. Some moderation systems for hate are easily fooled by minor changes in text or perform substantially worse on short length content or on certain types of media.[168]

- **Explainable**: Systems that have understandable decision-making processes and where decisions and outcomes are explained to end-users. Both Human- and AI-driven systems can involve black boxes and opaque processes, which may alienate and confuse users or may give undecipherable results.[169] Explainable systems tell the end users *why* their content has been classified and/or

[166] Ninareh Mehrabi et al., "A survey on bias and fairness in machine learning", *arXiv:1908.09635v2* (2019). Available at: https://arxiv.org/pdf/1908.09635.pdf.

[167] Thomas Davidson et al., "Racial Bias in Hate Speech and Abusive Language Detection Datasets", in *Proceedings of the Third Workshop on Abusive Language Online* (Florence: Association for Computational Linguistics), pp. 25-35 (2019). Available at: https://www.aclweb.org/anthology/W19-3504.;

Maarten Sap et al., "The Risk of Racial Bias in Hate Speech Detection", in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence: Association for Computational Linguistics), pp. 1668–1678 (2019). Available at: https://www.aclweb.org/anthology/P19-1163.

[168] Paul Röttger et al., "HateCheck: Functional Tests for Hate Speech Detection", *arXiv:2012.15606v1* (2020). Available at: https://arxiv.org/pdf/2012.15606.pdf.

[169] Bertie Vidgen et al., "Recalibrating classifiers for interpretable abusive content detection classifiers", *Proceedings of the Fourth Workshop of the Natural Language Processing and Computational Social Sciences*, pp. 132-133 (2020). Available at: https://www.aclweb.org/anthology/2020.nlpcss-1.14.pdf.

moderated in a particular way, such as highlighting which features of the content led to the outcome (i.e., 'Our AI identified that your use of the term 'f*g' is likely to be hateful').

- **Scalable**: Systems that can handle large volumes of data without sacrificing the other four desirable traits listed here (performance, fairness, robustness and explainability). Scalability includes (a) the system's cost-effectiveness and (b) its environmental and social impact.

Online platforms, including VSPs, can be motivated by a range of factors to improve how they tackle online hate, including the enforcement of legal and regulatory obligations and pressure from activist campaigns (see Section 1.1). In practice, most platforms do not operate entirely independently and the role of other service providers in the online technology ecosystem is an often-overlooked source of influence. Gillespie describes the process of **'stacked' content moderation** whereby "moderation decisions get made all up and down the infrastructural stack of services" (p. 6).[170] For instance, most platforms rely on cloud servers to host their services and on app stores to let users find and access them. These are two potential leverage points which can be used to demand higher moderation standards. This can be effective in challenging the worst purveyors of hate as "speakers banned across the many stacked and overlapping services will experience deplatforming to a much deeper degree" (p. 7) but at the same time it can worsen the opaqueness of content moderation and centralise power in the hands of a few companies – who may not always use such power consistently or fairly. Further, many companies are unwilling to adopt an interventionist stance and prefer instead to avoid making potentially contentious decisions. Nonetheless, several recent examples attest to the potential effectiveness of stacked moderation. Following the El Paso shooting in 2019 when 23 people were killed, the cloud service provider CloudFlare terminated

[170] Tarleton Gillespie, "Looking beyond Facebook: moderation everywhere" in Tarleton Gillespie et al. "Expanding the Debate about Content Moderation: Scholarly Research Agendas for the Coming Policy Debates", pp. 4-7.

service from 4chan, a largely unmoderated platform which has been closely associated with hate and extremism.[171] Following the Capitol riot in January 2020, Parler was removed from the Google Play and Apple stores and denied service by AWS. This shut down the 'free speech' platform immediately.[172]

To help understand how a content moderation system for hateful content could be designed, we distil the core requirements into four activities (shown in Figure 2). We use these four activities to structure the remainder of this section of the report (Sections 3.2 to 3.5) before discussing other important issues which must be considered in content moderation (Section 3.6) and ways of addressing online hate beyond content moderation (Section 3.7).

1. **Characterise online hate.** Provide a definition, typology and framework as needed.
2. Deploy strategies to **Identify online hate**.
3. **Handle online hate** through a proportionate response.
4. **Enable user complaints** through a robust and accessible review procedure.

Finally, we note one key limitation of content moderation; it is only one kind of intervention that will not replace considering the much greater problem of the drivers and causes of online hate. The United Nations' 2019 Strategy and Plan of Action on Hate Crime outlines that "addressing hate speech, therefore, requires a coordinated response that tackles the root causes and drivers of hate speech, as well as its impact on victims and societies more broadly."[173] This is increasingly well-recognised across the criminal

[171] Catherine Shu, "Cloudflare will stop service to 8chan, which CEO Matthew Prince describes as a 'cesspool of hate'", *Tech Crunch*, 5 August 2019. Available at: https://techcrunch.com/2019/08/04/cloudflare-will-stop-service-to-8chan-which-ceo-matthew-prince-describes-as-a-cesspool-of-hate/.

[172] Sarah Perez, "Following riots alternative social apps and private messengers top the app stores", *Tech Crunch*, 11 January 2021. Available at: https://techcrunch.com/2021/01/11/following-riots-alternative-social-apps-and-private-messengers-top-the-app-stores/.

[173] United Nations, *United Nations Strategy and Plan of Action on Hate Speech.*

justice system. For instance, the Mayor of London's Violence Reduction Unit was set up in 2018 to tackle violent crime by "identifying the root causes and delivering early interventions to help prevent its spread".[174] However, addressing the root causes of online hate is a difficult task and there is a relative lack of evidence apropos what is effective. Thus, whilst we strongly encourage more research into understanding the causes of online hate, content moderation is a more tractable way of addressing it in the immediate instance.

3.2 Step one: Characterise online hate

Online platforms typically characterise online hate in their Terms of Service and/or Community Guidelines.[175]As discussed above, creating a definition and/or typology of online hate is a difficult task and the depth and detail of definitions will vary across platforms. See Table 1 in the Appendix for a list of relevant definitions of online hate provided by popular platforms in the UK.[176] We identify three aspects for characterising online hate:

1. **Define online hate**. Definitions are generally short and can be easily understood. In this report we have already identified two exiting definitions: one from The Alan Turing Institute in Part 2 (and the Summary) and one in the AVMSD in Part 1, which is based on the EU's Council Framework Decision 2008/913/JHA.

2. **Construct a typology**. A typology is a classification framework which can distinguish between different varieties of a phenomenon. In Part 2 of this report we offer a two-part typology for online hate, which can be used to categorise

[174] Mayor of London, "How the Violence Reduction Unit is tackling the root causes of crime", 17 July 2019. Available at: https://www.london.gov.uk/city-hall-blog/how-violence-reduction-unit-tackling-root-causes-crime.

[175] Also referred to as 'Community Standards' and 'Transparency Rules'.

[176] Ofcom, "Online Nation 2020 Report" (London: Ofcom, 2020). Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0027/196407/online-nation-2020-report.pdf

online hate based on its *substance* and *articulation*. A typology may not always be needed, depending on platforms' requirements.

3. **Outline guidelines**: Guidelines provide detailed explanations of the line between different types of content, including where the line between hate and non-hate falls. They typically include examples (both exemplars and edge cases), rationales and principles.

Outlining the guidelines is often where the greatest disagreements appear as even seemingly 'neutral' or 'minimal' accounts may become embroiled in contentious and complex debates when they are applied to real content. Further, it is likely that different users, workers and external stakeholders may interpret the same guidelines in different ways, which can lead to divergent understanding of what content is hateful. 'Tricky' issues that guidelines are likely to need to consider include:

1. **Self-hatred**: Individuals may criticize a group to which they belong, which is an important part of civic discourse. In some cases, such criticisms will appear similar to hatred made against their community by others and in other cases it will be genuinely hateful, especially for people who have renounced their group.

2. **Truth and validity**: Hate can be expressed through pseudo factual statements which derogate a group (e.g., "You are X times more likely to be robbed by a Y than a Z"). These statements often rely on either low quality evidence or have been taken out of often taken out of context. In some cases, the 'evidence' will actually be false but nonetheless the statement will appear truthful. Such attacks can be deeply harmful but hate that is expressed with evidential support can be difficult to tackle without accusations of bias and censorship.

3. **Humour**: Jokes can be used to express hateful ideas, often exploiting the bigotries of the audience. It can be difficult to distinguish hateful jokes from jokes which

undermine, expose or satirise hatred.[177] Alt-right Internet subcultures are known for publishing offensive, trolling and tongue-in-cheek hateful content, often referred to as 'shit-posting'.[178]

4. **Intention**: Some accounts of online hate focus primarily on the intention of the speaker and whether they mean to inflict harm on a victim; it has been proposed that where there is no evidence of malicious intention then content should not be considered hateful.[179] However, this is problematic given that harm may still be inflicted, and discerning intention online is difficult.[180]

5. **Satire**: Content can use satire and irony to undermine and challenge hate. However, in so doing it may appear genuinely hateful. If viewed by unaware audiences, it could have similar effects to intended hate. This is a sensitive issue and some have questioned whether being satirical about prejudicial itself belies a lack of concern for the experiences of those who are targeted by prejudice.

To illustrate the difficulties of applying terms and conditions for online hate, Box 7 shows Facebook's three-tier approach to categorising hate speech (as of December 2020). The platform's policies have evolved substantially over the past decade and increasingly show sustained engagement with 'tricky' issues, as well as the challenges of balancing

[177] Bertie Vidgen et al., "Challenges and frontiers in abusive content detection", pp. 80-93;
Ji Hoon Park et al., "Naturalizing Racial Differences Through Comedy: Asian, Black, and White Views on Racial Stereotypes in Rush Hour 2", *Journal of Communication,* 56: 1, pp. 157-177 (2006). Available at: https://doi.org/10.1111/j.1460-2466.2006.00008.x..

[178] Luke Munn, "Alt-right pipeline: Individual journeys to extremism online", *First Monday*, 24: 6 (2019). Available at: https://doi.org/10.5210/fm.v24i6.10108.

[179] Anne Weber, *Manual on hate speech* (Strasbourg: Council of Europe Publishing, 2009);
Catherin O'Regan, "Hate Speech Online: an (intractable) contemporary challenge?", *Current Legal Problems*, 71: 1, pp. 403-429 (2018) Available at: https://doi.org/10.1093/clp/cuy012;
Naganna Chetty and Sreejith Alathur, "Hate speech review in the context of online social networks", *Aggression and Violent Behaviour*, 40, pp. 108-118 (2018) Available at: https://doi.org/10.1016/j.avb.2018.05.003.

[180] Bertie Vidgen et al., "Challenges and frontiers in abusive content detection", pp. 80-93.

protection from hate with the right to freedom of expression. Nonetheless, Facebook's approach to online hate has attracted substantial criticism for being reactive and inconsistent.[181] For instance, during 2020 Facebook was criticised for allowing content that denies the Holocaust, with accusations that it sometimes promotes such content through its recommender system algorithms.[182] Subsequently, in October 2020 its hate speech policy was updated to "prohibit any content that denies or distorts the Holocaust".[183] Equally, following a backlash against Instagram, which is owned by Facebook, for its delay in responding to British rapper and DJ Wiley's anti-Semitic content made in July 2020, Facebook also explicitly banned anti-Semitic conspiracy theories about Jewish people "controlling the world".[184]

Box 7: Facebook's characterisation of online hate

Facebook publishes various reports detailing its process for removing hateful content, as well as how these policies and processes are updated.[185] Its Community Guidelines outline its definition of hate speech:

[181] Alex Hern and Julia Carrie Wong, "Facebook Employees Hold Virtual Walkout Over Mark Zuckerberg's Refusal To Act Against Trump", *The Guardian*, 1 June 2020. Available at: https://www.theguardian.com/technology/2020/jun/01/facebook-workers-rebel-mark-zuckerberg-donald-trump.

[182] Jakob Guhl and Jacob Davey, "Hosting the 'Holohoax': A Snapshot of Holocaust Denial Across Social Media" (London: Institute for Strategic Dialogue, 2020). Available at: https://www.isdglobal.org/wp-content/uploads/2020/08/Hosting-the-Holohoax.pdf.

[183] Monika Bickert, "Removing Holocaust Denial Content" (San Francisco: Facebook, 12 October 2020). Available at: https://about.fb.com/news/2020/10/removing-holocaust-denial-content/.

[184] Alex Hern, "Facebook And Instagram Ban Antisemitic Conspiracy Theories And Blackface", *The Guardian*, 12 August 2020. Available at: https://www.theguardian.com/technology/2020/aug/12/facebook-and-instagram-ban-antisemitic-conspiracy-theories-and-blackface.

[185] Richard Allan, "Hard Questions: Who Should Decide What Is Hate Speech in an Online Global Community?", *Facebook* (San Francisco: Facebook, 27 June 2017). Available at: https://about.fb.com/news/2017/06/hard-questions-hate-speech/.

“a direct attack on people based on what we call protected characteristics — race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability. We protect against attacks on the basis of age when age is paired with another protected characteristic, and also provide certain protections for immigration status.

We define attack as violent or dehumanising speech, harmful stereotypes, statements of inferiority, or calls for exclusion or segregation”[186]

Facebook uses a triaging system to separate online hate into three tiers. Key parts of the tiers are shown below. [187]

Tier 1	Tier 2	Tier 3
● Violent speech or support in written or visual form ● Dehumanising speech or imagery. ● Mocking the concept, events or victims of hate crimes even if no real person is depicted in an image	● Generalisations that state inferiority (in written or visual form) in the following ways: Physical deficiencies, Mental deficiencies and Moral deficiencies. ● Other statements of inferiority (including expressions about	● Calls for segregation ● Explicit Exclusion which includes but is not limited to "expel" or "not allowed". ● Political Exclusion defined as denial of right to political participation. ● Economic Exclusion defined as denial of

[186] Facebook, “Community Standards: Hate Speech”. Available at: https://www.facebook.com/communitystandards/recentupdates/hate_speech/. Last accessed on 4 December 2020.

[187] Ibid.

● Designated dehumanising comparisons, generalizations, or behavioural statements (in written or visual form).	being less adequate) ● Expressions of contempt (including expressions that a protected characteristic shouldn't exist) ● Expressions of dismissal ● Expressions of disgust ● Cursing (which includes referring to the target as genitalia or anus).	access to economic entitlements and limiting participation in the labour market. ● Social Exclusion defined as including but not limited to denial of opportunity to gain access to spaces (incl. online) and social services.

3.3 Step two: Identify online hate

Once online hate has been characterised, ways of identifying it need to be implemented. This will vary across platforms, depending on their expertise, infrastructure and budget. Broadly, three planks form the basis of most content moderation processes for identifying online hate: User reports, AI, and human review. One illustrative example of how hate could be identified is given in Figure 2, showing how the three planks can be combined. Note that this example is not based on any single platform and that the three planks can also be combined in other ways.

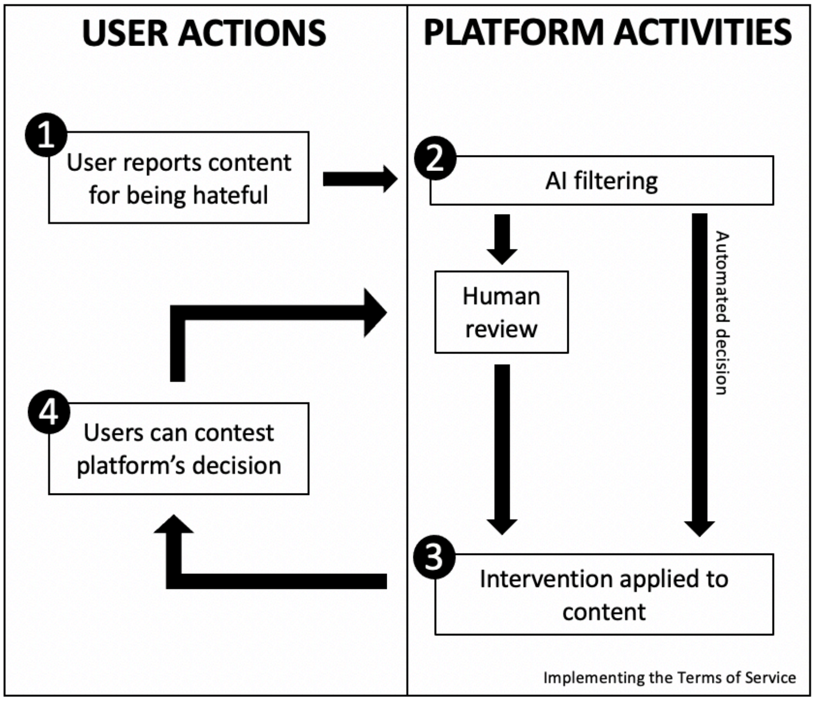
USER ACTIONS
PLATFORM ACTIVITIES
1
User reports content for being hateful
2
AI filtering
Human review
Automated decision
3
Intervention applied to content
4
Users can contest platform's decision
Implementing the Terms of Service

Figure 2: An example content moderation system, combining user reports, AI and human review.

In Figure 2, a user reports content which they think is hateful and violates the Terms of Service. Typically, users can click on content and then select an option to report it to the platform, although the ease of reporting varies substantially across platforms.[188] Then, AI is used to filter the content. Some content will have an intervention applied automatically (see the next Section) whereas other content will be sent for human review. After the intervention has been applied to the content, users can contest the platform's decision (see Section 3.5).

Another way of identifying harmful content is to enable users to self-flag their content at the point of upload. This is a well-established practice with content which might be

[188] Paul M. Barrett, *Who Moderates the Social Media Giants? A Call to End Outsourcing* (New York: NYU Stern Center for Business and Human Rights, 2020). Available at: https://issuu.com/nyusterncenterforbusinessandhumanri/docs/nyu_content_moderation_report_final_version

harmful to minors, such as pornographic content and violent material. To our knowledge, it has not been used as a strategy to tackle online hate specifically – which is understandable given that it is unlikely that many people would tag their own content as 'hateful'. An approach used by BitChute is to let users select a "content sensitivity" rating when uploading content. By default, content is marked as "Normal" (suitable for people over the age of 16) but users can also rate their content as "Not Safe For Work (NSFW)" and "Not Safe For Life (NSFL)". The NSFW setting is for content that "is not safe for viewing in the workplace, or similar environments", such as videos containing nudity, moderate violence, drug use and/or discriminatory language. NSFL is the highest level of sensitivity "as it does not matter where you view the material; many if not most people will find this content upsetting". BitChute does not rely solely on self-tagging and its Community Guidelines state that other content moderation processes are also enforced.[189]

3.3.1 Human review of hateful content

Human moderators can be used to check and review any content that has been flagged and make a decision about how it should be handled. Well-trained human moderators can have expert knowledge of a particular online setting and subject matter, understand the cultural norms and linguistic tropes, and can flexibly react to new developments and policies. However, human-led moderation can also be time-consuming, expensive and can inflict social and psychological harm on the moderators. It may lead to inconsistent results if moderators are inadequately trained or are not given sufficient time to review each bit of content. Concerningly, the personal costs imposed on moderators are often not fully recognised; Roberts argues that while the work of content moderators is essential it is also "seemingly paradoxically, invisible" (p.1).[190] She claims that large social

[189] BitChute, "Community Guidelines: Content Sensitivity". Available at: https://support.bitchute.com/policy/guidelines/#content-sensitivity. Last accessed on 4 December 2020.

[190] Sarah T. Roberts, *Behind the Screen: Content Moderation in the Shadows of Social Media* (New Haven; London: Yale University Press, 2019).

media companies enact “a series of distancing moves designed to create a plausible deniability to limit their responsibility for workplace harm, particularly when such harm may take time to show up” (p.127).[191]

Numerous investigations suggest that moderators are often underpaid and receive inadequate care and support. They have reported working long hours in pressurised working environments where they are expected to achieve high daily targets and can be fired for making errors.[192] For example, it was reported in January 2020 that one of the third-party companies which provide moderators for platforms such as Facebook and YouTube had not provided counselling or medical care for its employees. Instead, employees had only been given ‘wellness’ programmes, which were criticised for being inadequate.[193] Some moderators have developed post-traumatic stress disorders resulting from continued exposure to hateful content and some have reported embracing the extreme or fringe viewpoints that they are moderating.[194] In 2019, an investigation by Bloomberg found that outsourced content moderators at one site had responded to their difficult working environment by consuming alcohol and marijuana.[195]

Moderation is often outsourced to lower cost sites across the world, with major hubs in countries such as the Philippines, India, and Ireland. In a high-profile report published in 2020, Barrett argues that such international outsourcing not only exploits workers, it also

[191] Ibid.

[192] Casey Newton, "The Secret Lives Of Facebook Moderators In America", *The Verge*, 25 February 2019. Available at: https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona.

[193] Casey Newton, “YouTube Moderators Are Being Forced To Sign A Statement Acknowledging That The Job Can Give Them PTSD”, *The Verge*, 24 January 2020. Available at: https://www.theverge.com/2020/1/24/21075830/youtube-moderators-ptsd-accenture-statement-lawsuits-mental-health.

[194] Ibid, "The Secret Lives Of Facebook Moderators In America".

[195] Joshua Brustein, “Facebook Grappling With Employee Anger Over Moderator Conditions”, *Bloomberg*, 25 February 2019. Available at: https://www.bloomberg.com/news/articles/2019-02-25/facebook-grappling-with-employee-anger-over-moderator-conditions. Last accessed on 4 December 2020.

leads to reduced performance in a critical function.[196] This is because online hate is often nuanced, contextual and requires understanding of cultural and social factors. Workers from a different culture who are non-native speakers with inadequate training are unlikely to make appropriate moderation decisions. This reflects a well-established principle in academic research into online hate, where experts such as Waseem have long argued that "hate speech is hard to [assess] without intimate knowledge of hate speech."[197]

3.3.2 The use of AI to flag online hate

AI has the potential to automatically flag whether or not content is hateful[198] and The Alan Turing Institute has conducted numerous research projects in this area, developing classification tools for different types of online hate, as well as investigating how such tools are evaluated and implemented.[199] Advanced computational models learn representations of hateful content from large training datasets, which they then use to automatically label new content they are presented with.[200] Recent advances in the algorithms and models which underpin AI have led to huge improvements in their performance at detecting and classifying online hate.[201]For instance, in its Q3 2020

[196] Paul. M Barrett, *Who Moderates the Social Media Giants? A Call to End Outsourcing.*

[197] Zeerak Waseem, "Are you racist or am I seeing things? Annotator influence on hate speech detection on Twitter", *Proceedings of the 2016 EMNLP Workshop on Natural Language Processing and Computational Social Science*, pp. 138-142 (2016). Available at: https://www.aclweb.org/anthology/W16-5618

[198] Cambridge Consultants, *Use of AI in online content moderation* (London: Ofcom, 2019). Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf.

[199] The Alan Turing Institute, "Hate Speech: Measures and Counter-Measures". Available at: https://www.turing.ac.uk/research/research-projects/hate-speech-measures-and-counter-measures. Last accessed on 4 December 2020.

[200] The Royal Society*, Machine learning: the power and promise of computers that learn by example* (London: The Royal Society, 2017). Available at: https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf?la=en-GB&hash=B4BA640A1B3EFB81CE4F79D70B6BC234.

[201] Bertie Vidgen et al., "Challenges and frontiers in abusive content detection", pp. 80-93; Dynabench, "Dynabench". Available at: https://dynabench.org/about. Last accessed on 4 December 2020.

transparency reporting, Facebook stated that over 95% of all hate speech had been identified using AI.[202]

The great promise of AI lies in (a) reducing the burden placed on human content moderators, (b) increasing the speed with which online hate is tackled and (c) minimising human biases and inconsistencies in the moderation process. Yet even state-of-the-art AI technologies have numerous limitations and their unmanaged use can lead to undesirable outcomes. For example, in May 2020, YouTube admitted that it had automatically deleted a large number of comments containing certain phrases critical of the Chinese Communist Party.[203] The content was in no way hateful, offensive or spam but, instead, legitimate political speech. The platform described the mistake as an "error with our enforcement systems".[204]

Errors in AI can be particularly severe, depending on the type, articulation and media of content. For example, video processing is a challenging area of machine learning and substantially less research has investigated how to train and improve automated systems for hateful video flagging compared with text-based content.[205] Most detection systems are prone to being circumvented through adversarial attacks in which minor features of content are subtly changed (in a way that is imperceptible to the human eye) and are then misclassified. This was shown during the Christchurch attack in 2019 when malicious

[202] Facebook, *Community Standards Enforcement Report Q3 2020.*

[203] James Vincent, "YouTube is deleting comments with two phrases that insult China's Communist Party", *The Verge*, 26 May 26 2020. Available at: https://www.theverge.com/2020/5/26/21270290/youtube-deleting-comments-censorship-chinese-communist-party-ccp.

[204] James Vincent, "YouTube says China-linked comment deletions weren't caused by outside parties", *The Verge*, 28 May 2020. Available at: https://www.theverge.com/2020/5/28/21272983/youtube-deleting-comments-chinese-communist-censorship-explanation.

[205] Ashish Sureka et al., "Mining YouTube to Discover Extremist and Hidden Communities", *Lecture Notes In Computer Science: Asia Information Retrieval Symposium*, 6458, pp. 13-24 (2010). Available at: https://link.springer.com/chapter/10.1007/978-3-642-17187-1_2.;
Swati Argawal and Ashish Sureka, "A focused crawler for mining hate and extremism promoting videos on YouTube", *Proceedings of the 25th ACM on Hypertext and Social Media*, pp. 294-296 (2014). Available at: https://doi.org/10.1145/2631775.2631776

actors on Facebook repeatedly uploaded first-hand videos of the attack with almost-unnoticeable adjustments; 1.2 million videos of the livestreamed attack were blocked from being uploaded through AI but a further 300,000 were not immediately flagged due to these minor adjustments.[206] At the same time, there is clear potential for AI to improve how hateful videos are detected. A Home Office sponsored project in the UK developed an algorithm to detect extreme pro-ISIS videos; it achieved 94% recall and had a false positive rate of only 0.005%.[207]

AI is imperfect and many AI-driven systems will not perform well against all of the design features we identified in Section 3.1. To help clarify the key weaknesses of even State of the Art AI (as of December 2020) we identify ten challenges for hateful content detection. That said, it should be noted that every AI system is different, and advances are constantly being made in how AI is trained, evaluated and implemented.

1. AI often lacks understanding of the wider social and historical context.
2. AI typically lacks understanding of the speaker's identity and their previous online activity; giving it such information to learn from may involve unacceptable levels of data harvesting.
3. AI struggles with satire and other complex forms of expression.
4. AI can be highly biased and may have different levels of accuracy on different social groups. This can perpetuate social unfairness and injustice.
5. AI can struggle with content of longer length, especially if the hate is expressed through multiple sentences and requires understanding of how they relate to each other.
6. AI can fail to identify hateful content which is part of a *conversational dynamic*, such as content which expresses support for another user's hate.

[206] John Gallacher, "Automated Detection of Terrorist and Extremist Content" in Bharath Ganesh and Jonathan Bright (eds.) *Extreme Digital Speech* (London: Vox-POL, 2019). Available at: https://www.voxpol.eu/download/vox-pol_publication/DCUJ770-VOX-Extreme-Digital-Speech.pdf

[207] BBC News, "UK unveils extremism blocking tool", 13 February 2018. Available at: https://www.bbc.co.uk/news/technology-43037899.

7. AI generally performs worse on multimedia and multimodal content such as videos, as well as memes, images and audio.
8. AI can be difficult to update over time and models may fast become outdated as online hate changes.
9. AI can lack robustness to small changes in content.
10. AI may fail to identify content which uses coded language, whereby seemingly neutral terms are used as replacements for racial slurs and/or content is intentionally obfuscated to avoid detection.[208] These challenges mean that AI should supplement rather than supplant humans in the content moderation process. A hybrid approach to content moderation, in which human moderators are integrated with AI, is likely to be most effective – and AI should not be used without a 'human in the loop' who has ultimate oversight of its decision-making.

3.4 Step three: Handle online hate

Once online hate has been identified, it needs to be handled. Policy-making and civic discourses often focus overwhelmingly on the implications of *banning* users and their content.[209] Yet in practice platforms use different strategies to tackle online hate

[208] Tommi Gröndahl et al., "All You Need is "Love": Evading Hate Speech Detection", in *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security* (New York: Association for Computing Machinery), pp. 2–12 (2018). Available at: https://doi.org/10.1145/3270101.3270103.;
Rijul Magu et al., "Detecting the Hate Code on Social Media," in *Proceedings of the Eleventh International AAAI Conference on Web and Social Media* (California: Association for the Advancement of Artificial Intelligence) (2017). Available at: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15604.

[209] Raluca Balica, "The Criminalisation of Online Hate Speech: It's Complicated", *Contemporary Readings in Law and Social Justice*, 2:1, pp. 184-190 (2017). Available at: https://www.ceeol.com/search/article-detail?id=589426.; Stefanie Ullmann and Marcus Tomalin, "Quarantining online hate speech: technical and ethical perspectives", *Ethics and Information Technology*, 22:1, pp. 69-80 (2020). Available at: https://doi.org/10.1007/s10676-019-09516-z;
Jessica Henry, "Beyond free speech: novel approaches to hate on the Internet in the United States", I*nformation & Communications Technology Law,* 18: 2, pp. 235-251 (2009). Available at: 10.1080/13600830902808127.;

depending on the nature of the content. Banning may also not be that effective; Gagliardone et al. contend that, "[e]ven when content is removed, it may find expression elsewhere, possibly on the same platform under a different name or on different online spaces." (p.14) [210] Different strategies for handling online hate impose different levels of **friction**, which we define as:

> The degree of resistance that content encounters in order to be published and found, seen, shared and engaged with by audiences.

Banning users imposes the greatest friction as it usually means that all of their historical content is removed and they are unable to post any new content. At the other end of the moderation spectrum are interventions such as constraining how many times users can share content, which impose a far lower level of friction. The use of different interventions is important for ensuring hate is dealt with proportionately. Table 2 shows 14 different strategies for handling online hate, including who is affected by each strategy. We group the strategies into four buckets:

1. **Hosting constraints**: Is the content and/or user allowed on the platform?
2. **Viewing constraints**: Are users able to view the content without restrictions?
3. **Searching constraints**: Are users able to find the content easily?
4. **Engagement constraints**: Are users able to interact with the content, such as liking/sharing/commenting?

In practice, platforms often combine these strategies; engagement constraints (such as banning comments/likes) might also be implemented with viewing constraints (e.g., requiring explicit consent to view) and search constraints (e.g., stopping paid-for

Cherian George. "Hate Speech Law and Policy" in P.H. Ang and R. Mansell (eds.) *The International Encyclopedia of Digital Communication and Society* (New Jersey: Wiley-Blackwell, 2015). Available at: https://doi.org/10.1002/9781118767771.wbiedcs139.

[210] Iginio Gagliardone et al., *Countering online hate speech.*

promotions). The degree of friction which strategies actually create will also depend upon how they are implemented.

For simplicity, we have focused on actions which are implemented by platforms, although in practice some strategies will be implemented by third-party paid moderators, such as managers of a company's social media presence, or by community-nominated page owners and forum managers.

Constraints	**Moderation strategy**	**Description**	**Who does the strategy affect?**
Hosting constraints	Ban users	Users are banned permanently from a platform. They are usually also banned from creating new accounts. This is the most severe option and is usually only used for spam and bot accounts and repeated offenders of Terms of Service. Banning users from creating new accounts is technically difficult, especially with the widespread availability of IP-masking technology.[211]	The hateful user[212]

[211] James Titcomb, "Twitter blocks banned users from creating new accounts", *The Telegraph*, 7 February 2017. Available at: https://www.telegraph.co.uk/technology/2017/02/07/twitter-blocks-banned-users-creating-new-accounts/.

[212] 'Hateful user' refers to a user who is hateful towards others.

	Suspend users	Users are temporarily banned. The length of time that users are banned for usually increases with repeated violations of the Terms of Service and/or the severity of the violation.	The hateful user
	Remove content (permanently)	Content is removed permanently.	The hateful user
	Remove content (temporarily)	Content is removed temporarily. During the suspension period it may be reviewed and subsequently reinstated or taken down permanently.	The hateful user
	Prompt users	Before content is posted, the user is given a prompt about its possible impact.[213]	The hateful user
Viewing constraints	Require explicit consent to view content	Users are required to give their explicit consent before they can view it. This is typically implemented through an interstitial page on the VSP.	Viewers of the hateful content

[213] The BBC's Own It App is an example of this strategy.

	Show content with a warning	A warning is attached to content, identifying to users that it is harmful, toxic or otherwise contentious. The content can still be viewed by the user. Warnings may be displayed as text or with icons (such as fact-checking logos).	Viewers of the hateful content
	Showing content with a competing viewepoint/counterspeech	Content is shown next to competing content, either from trusted flaggers or rival communities. To our knowledge, no platforms use this strategy to tackle online hate. The only equivalent tools that we are aware of have been developed by third parties for online news.[214]	Viewers of the hateful content
Search constraints	Make content unsearchable	Content is removed from any search results provided on the platform, minimising the likelihood that new users will see the new content and ensuring there will be little 'accidental' engagement.	Viewers of the hateful content

[214] Left/Right News, "Look Left. Look Right. Think Straight." (2020). Available at: https://leftright.news. Last accessed on 4 December 2020.;
AllSides, "Don't be fooled by media bias and fake news." (2020). Available at: https://www.allsides.com/unbiased-balanced-news. Last accessed on 4 December 2020.;
Read Across The Aisle, "Read Across the Aisle" (2020). Available at: http://www.readacrosstheaisle.com. Last accessed on 4 December 2020.

	Algorithmically downvote content	Content is downvoted or removed from users' timelines, feeds and recommendations, reducing the likelihood that other users will view it.	Viewers of the hateful content
	Stop paid-for-promotions of content	Users are stopped from paying to promote content, minimising their ability to engage with new audiences.	The hateful user
	De-monetise content	Users are unable to receive income from views of content, such as being banned from YouTube's content-creator payment model.	The hateful user
Engagement constraints	Constrain how many times each user can share content	Users are constrained in how many times they can share and repost content, limiting the spread of harmful 'viral' content.	Viewers of the hateful content
	Ban commenting on and/or liking content	Users are unable to comment on or like content.	Viewers of the hateful content

Table 2: Different ways of handling online hate.

Friction can create trade-offs, however. Limiting users' ability to post, promote, engage with, view and share content can constrain freedom of expression, lead to unintended negative consequences (such as silencing certain political voices), reduce engagement on mainstream platforms (potentially motivating users to migrate towards less-well-regulated niche/alternative spaces) and may involve heightened data ingestion to track users' behaviour and implement the moderation strategies. Two substantial secondary risks are also posed. First, the implementation of any content moderation strategy is by nature imperfect (see the analysis above). As a result, some non-hateful content will be miscategorised as hateful; strategies which impose too much friction will also be (mis)applied to this content. Second, all strategies risk creating 'chilling effects', whereby free discourse is constrained by people's awareness that they *could* be penalised. Creating too much friction could contribute to this problem as overly punitive responses to minor infractions will heighten the sense that free expression is not allowed.

Which moderation strategy is used by services will depend on many factors, including financial cost, how long they take to implement, technical feasibility and the company's values and culture. Deciding which strategy should be applied is clearly an imperfect science but should be approached robustly and transparently, with rationales given for *why* the different strategies have been selected. In principle, *the more harm that is likely to be inflicted, the more friction can be justified*. One reasonable approach would be to triage the amount of friction that is imposed. Content that is highly harmful, and likely to be illegal, would face the greatest friction, such as user bans, whilst content which is less harmful would face less friction. This is shown in Figure 3.

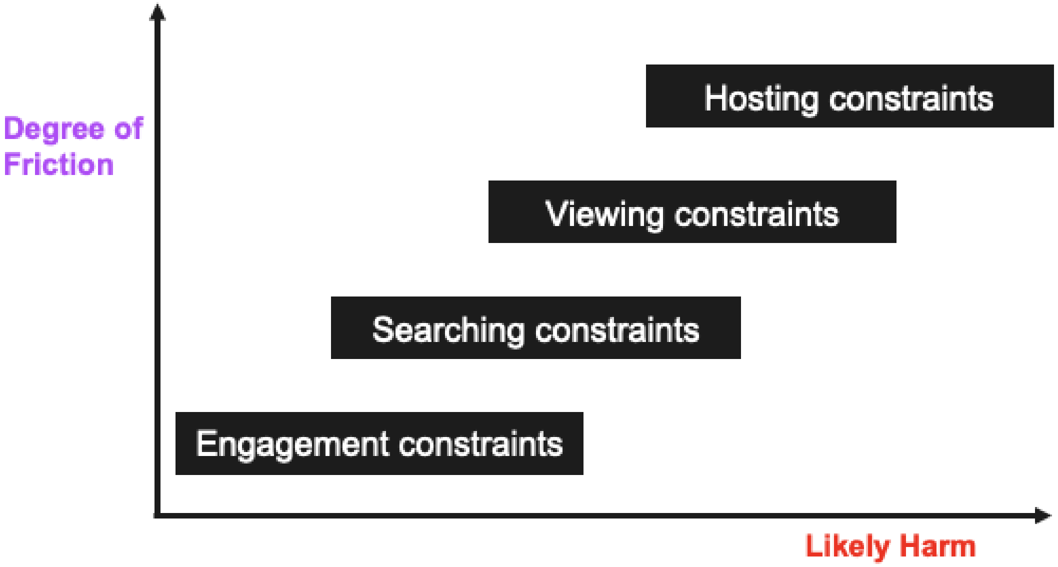

Figure 3: Degree of friction for different moderation strategies versus the likely harm that is inflicted.

3.5 Step four: Enable users to appeal decisions

All content moderation systems will make mistakes and involve making decisions about contentious issues. It is therefore important that users can challenge the decisions they are subjected to and request that any moderation is overturned. We draw attention to three key areas in user appeals: (1) the information users are given, (2) whether users are involved in the content moderation process and (3) the speed at which users' content is moderated. To illustrate an example moderation system, YouTube's process is shown in Box 8.

Box 8: YouTube user referral system for violating content

As of December 2020, YouTube enforces a 'strike' system for content which violates its Terms of Use:

> "When we remove your content for a Community Guidelines violation, you may be issued a strike. Strikes are issued when content on YouTube is flagged for

> review, either by members of the YouTube community or our smart detection technology, and our review teams decide that it does not follow our Community Guidelines. If your channel gets a strike, you'll get an email, notifications on mobile and desktop, and an alert in your channel settings the next time you sign in to YouTube." [215]

When notified, the user is given information on:

- What content was removed;
- Which policies it violated (e.g., adult content or violence);
- How it affects the channel;
- What the user can do next.

The platform provides users with the option to appeal the removal of their content, but they can only appeal 30 days after the warning or strike was issued. After the appeal is sent, one of the following outcomes is taken by YouTube:

- "If we find that your content followed our Community Guidelines, we'll reinstate it and remove the strike from your channel. If you appeal a warning and the appeal is granted, the next offense will be a warning.
- If we find your content followed our Community Guidelines, but isn't appropriate for all audiences, we'll apply an age-restriction. If it's a video, it won't be visible to users who are signed out, are under 18 years of age, or have Restricted Mode turned on. If it's a custom thumbnail, it will be removed.
- If we find that your content was in violation of our Community Guidelines, the strike will stay and the video will remain down from the site. There's no additional penalty for appeals that are rejected."

[215] YouTube, "Appeal Community Guidelines actions". Available at: https://support.google.com/youtube/answer/185111. Last accessed on 4 December 2020.

Google's Transparency Report provides data on all video removals on YouTube (for any reason). Between July and September 2020, 7.9 million videos were removed. 210,000 videos were appealed, of which 80,000 were reinstated.[216]

3.5.1 The information users are given

Platforms vary in the information given to users about how (and whether) their content is moderated. They also update their processes over time, meaning that any assessment is time-specific. For instance, TikTok only started to explain to users which specific policy their content has violated, as well as how the decision can be appealed, in October 2020.[217] Previously, it had only informed them that their content had "violated the company's guidelines."[218] In July 2019, Instagram introduced a notification system which told users whether their account was at risk of being disabled. [219] The update explains to users how removals are decided, which is based on the percentage of their content which violates the terms and conditions and the number of violations that users make.

Although the information provided to users about how their content is moderated is broadly improving, concerns have been raised that many platforms still do not keep users

[216] Google, "Appeals". Available at: https://transparencyreport.google.com/youtube-policy/appeals?hl=en_GB. Last accessed on 4 December 2020.

[217] TikTok, "Adding clarity to content removals", 22 October 2020. Available at: https://newsroom.tiktok.com/en-us/adding-clarity-to-content-removals. Last accessed on 4 December 2020.

[218] Sean Hollister, "TikTok will now tell you why it removed your video", *The Verge*, 22 October 2020. Available at: https://www.theverge.com/2020/10/22/21529497/tiktok-content-violation-which-policy-community-guidelines-update.

[219] Instagram, "Account Disable Policy Changes on Instagram". 18 July 2019. Available at: https://about.instagram.com/blog/announcements/account-disable-policy-changes-on-instagram. Last accessed on 4 December.

well-informed enough. In December 2020 the international human rights campaign group Article 19 called on the major platforms to change their user appeal process:

1. "Whenever companies take down user content or suspend an account, we want them to notify the user and clearly explain what content has been removed and why.

2. When notifying users of a take down or account suspension, we want companies to give users the opportunity to appeal the decision, using clear and simple language to tell them how to do this, and giving them the opportunity to discuss the matter with a person.

3. Finally, we want these companies to proactively publish much more detailed data on the numbers of complaints, content takedowns and appeals which have been made together with detail on the type of information that was removed and reinstated."[220]

Most platforms only give information to users when their content is taken down or their account is suspended or banned. Other interventions, particularly search constraints, are generally less well documented and may not even be made apparent to users. In some cases, platforms may make a deliberate decision to not give users full information. For instance, 'shadow banning' is where users are banned but do not know it and still believe they can post live content. In these cases, the fact that users are not aware that they are being subjected to any friction is a key *feature* of the intervention's design.[221] Thus, whilst more transparency is certainly welcome, further critical reflection is needed to

[220] Article 19, "#MissingVoices". Available at: https://www.article19.org/campaigns/missingvoices/. Last accessed on 4 December 2020.

[221] Preran Juneja et al., "Through the Looking Glass: Study of Transparency in Reddit's Moderation Practices", *Proceedings of the ACM on Human-Computer Interaction*, 4: 1 (2019). Available at: https://doi.org/10.1145/3375197

understand all of the implications given that more transparency could undermine the efficacy of interventions such as shadow bans.

3.5.2 Users' involvement in the content moderation process

Whether users are involved in the creation, implementation and management of the moderation process is another important concern, particularly given the speed at which hate can change and given that the social consensus about what should be permitted online can rapidly shift. Most platforms determine their own policies, often without sustained community engagement and little feedback from those who are affected by them. This can lead to inappropriate policies being developed and/or policies not being updated quickly enough. Part of the challenge is that platforms are usually unable to share all details of how they moderate content; there is a risk that such information could be weaponised by malicious actors who may exploit any weaknesses or gaps in the policies and/or their enforcement to 'game' them. Such issues need to be carefully weighed up before policies are made public and users are directly involved in their formation.

Some companies are exploring new content moderation policy structures to give users more voice. In October 2020, Facebook established the "Oversight Board" to help decide "significant and difficult" decisions in content moderation on both Facebook and Instagram. It is composed of independent members who rule on complex cases and has been funded with a grant of $130 million. [222] Users can refer a Facebook content moderation decision to the Oversight Board, which will also provide other advice to the platform through "nonbinding recommendations".[223] Appeals can only be made once users have already exhausted the existing appeals process and if they disagree with the

[222] Brent Harris, "Oversight Board to Start Hearing Cases", *Facebook*, 22 October 2020. Available at: https://about.fb.com/news/2020/10/oversight-board-to-start-hearing-cases/.

[223] Casey Newton, "Facebook's new Oversight Board is a wild new experiment in platform governance", *The Verge*, 23 October 2020. Available at: https://www.theverge.com/2020/10/23/21530524/facebooks-new-oversight-board-platform-governance.

platform’s decision.[224] Questions have already been raised about the Board, with some doubting whether it will be truly impartial.[225]

3.5.3 The speed at which users’ content is moderated

A final consideration is the speed at which content is moderated. This is particularly important for minimising the harm that online hate can inflict on users and for maintaining trust with them. As noted above, the EU Code of Conduct to Tackle Online Hate mandates platforms to review flagged content within 24 hours. Yet the time taken to moderate content is likely to vary over time as periods of high levels of hateful activity, such as following political events or terror attacks, may put a strain on available resources.[226] One strategy that could be used to address this is triaging, whereby platforms prioritise resources towards content that is likely to inflict the greatest harm (i.e., where it has a high level of hazard and/or is likely to exert substantial influence through the reach and status of the person who creates it). It remains important that all

[224] Facebook Oversight Board, “Announcing the Oversight Board's first cases and appointment of trustees”, December 2020. Available at: https://www.oversightboard.com/news/719406882003532-announcing-the-oversight-board-s-first-cases-and-appointment-of-trustees/.

[225] Elizabeth Culliford, “From hate speech to nudity, Facebook's oversight board picks its first cases”, *Reuters*, 1 December 2020. Available at: https://uk.reuters.com/article/facebook-oversight/from-hate-speech-to-nudity-facebooks-oversight-board-picks-its-first-cases-idINKBN28B50Y.;
Taylor Hatmaker, “Facebook’s controversial Oversight Board starts reviewing content moderation cases”, *TechCrunch*, 22 October 2020. Available at: https://techcrunch.com/2020/10/22/facebook-oversight-board-controversy/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAMWkyH86ZK2RbsAHdUsZL7LJKb0gz6iTeZwDyj1SC5Ye9HIqKwPdsIyyhds1yYq41fg5_zz8UL5AYtTd6t-512dUPPM-Nq8x1_IcW5wDREqoitVq84amL1GfFzuV8K3pXfJsXLf3LSmrGd6QZ3QcVXYV72h6kXBXSPNxq0PgPsEL.;
Casey Newton, “Facebook’s new Oversight Board is a wild new experiment in platform governance”.

[226] Bertie Vidgen and Taha Yasseri, “Four ways social media platforms could stop the spread of hateful content in aftermath of terror attacks”, *The Conversation*, 18 March 2020. Available at: https://theconversation.com/four-ways-social-media-platforms-could-stop-the-spread-of-hateful-content-in-aftermath-of-terror-attacks-113785.

content is addressed adequately and that no harmful content is either left online for an excessive amount of time or left in 'limbo' as it undergoes review.

3.6 Balancing online hate regulation with other concerns

The regulation of online hate raises fundamental ethical questions about the complex balancing act between protecting users from harm whilst ensuring others' rights are protected. In the remainder of this Section we explore the two most pressing issues in relation to online hate: freedom of expression and privacy.

3.6.1 Freedom of expression

Freedom of expression is a Human Right, protected by Article 10 of the European Convention on Human Rights:

> "Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers"[227] (p.11)

Freedom of expression is rightly protected as a fundamental part of civic debate and the main way in which ideas, values and beliefs are exchanged and critiqued in liberal democratic societies. In the 2019 legal case of *Miller v The College of Policing & Another,* the Judge made this argument explicitly, quoting George Orwell that, "If liberty means anything at all, it means the right to tell people what they do not want to hear."[228]

Freedom of expression is arguably the most high-profile and contentious issue in discussions of how to tackle online hate[229], and nearly all regulations and policies on hate

[227] European Court of Human Rights, *Guide on Article 10 of the European Convention on Human Rights: Freedom of Expression* (Brussels: Council of Europe, 2020). Available at: https://www.echr.coe.int/Documents/Guide_Art_10_ENG.pdf.

[228] Royal Courts of Justice, *The Queen on the application of Harry Miller v (1) The College of Policing and (2) The Chief Constable of Humberside CO/2507/2019* (London: Royal Courts of Justice, 2020). Available at: https://www.judiciary.uk/wp-content/uploads/2020/02/miller-v-college-of-police-judgment.pdf.

[229] Alexander Brown, *Models of Governance of Online Hate Speech*.

include provisions requiring the state to protect free expression, including the AVMSD.[230] Interventions are usually justified by a principled assessment of the *harm* that the content can inflict and whether this outweighs any potential constraint on freedom of expression. This discussion must be set in the context that moderation systems are imperfect and it is likely that some non-hateful content will be identified as hateful (and vice versa). This is particularly concerning given that some content which might initially seem harmful may actually be in the public interest. For example, videos of war atrocities may appear hateful but serve to draw attention to breaches of international law. Indeed, in response to the 2019 Online Harms White Paper, the civil liberties organisation Big Brother Watch argued that "To focus an enforcement framework that would affect vast swathes of modern communications on the undefined notion of "harm" would be to open the door to subjective and politicised censorship." (p.6).[231]

Draconian policies which are not supported by a proportionate balancing up of the harm caused by online hate with the benefits of freedom of expression could create more negative consequences than they mitigate. As Brown argues, "A truly responsible Internet platform is one that, on occasion and where appropriate, is willing to defend in the courts its decisions not to remove content, on the grounds of promoting and protecting the human right to freedom of expression." (p. 24).[232] We identify the following **six risks of excessive moderation of online hate**:

1. Free speech will be limited through the removal of content which is either non-hateful or expresses such a subtle and/or ambiguous form of prejudice as to be permissible.

[230] EU Legislation, *Directive (EU) 2018/1808 of the European Parliament and of the Council of 14 November 2018 amending Directive 2010/13/EU*.
[231] Big Brother Watch, *Big Brother Watch's response to the Online Harms White Paper Consultation* (London: Big Brother Watch, 2019). Available at: https://bigbrotherwatch.org.uk/wp-content/uploads/2020/02/Big-Brother-Watch-consultation-response-on-The-Online-Harms-White-Paper-July-2019.pdf.
[232] Alexander Brown, *Models of Governance of Online Hate Speech*.

2. Free speech could be limited through the subsequent ‘chilling effects’ of excessive content removal.[233]Politically sensitive perspectives and important emancipatory activism may be severely undermined. A 2019 report by the free speech organisation Index on Censorship found that following changes in YouTube’s hate speech policy, anti-racist activists had their channels removed or videos pulled down for containing slurs, while thousands of academic, journalistic and activist sites were removed as well.[234]

3. Users may be driven by the removal of content to migrate to smaller and less well-regulated platforms which may not come under the remit of the AVMSD or other regulation. They could be exposed to far more harmful and dangerous content in such spaces. This is discussed in Box 9.

4. Excessive moderation may further conspiracy theories about the management and governance of online spaces, including the motivations and aims of platforms. This could motivate users to adopt more dangerous beliefs and outlooks.

5. Platforms may lose revenue and activity from their users leaving or becoming active, which could threaten their long-term viability.

At the same time, it is important to acknowledge that *not* taking action on hate speech may constrain freedom of expressions just as much as being overly censorious. If hate is permitted then the groups which are targeted may no longer feel comfortable or safe taking part in public discourse and adopting prominent public positions; this may be more harmful for democratic debate than the loss of potential purveyors of online hate.

[233] Jonathon W. Penney, “Internet surveillance, regulation, and chilling effects online: a comparative case study”, *Internet Policy Review*, 6: 2, pp. 1-39 (2017). Available at: http://dx.doi.org/10.14763/2017.2.692.
[234] Index on Censorship, *Index on Censorship submission to Online Harms White Paper consultation* (London: Index on Censorship, 2019). Available at: https://www.indexoncensorship.org/wp-content/uploads/2019/07/Online-Harms-Consultation-Response-Index-on-Censorship.pdf.

Box 9: Deplatforming and user migration to other platforms

Online users who promote hateful narratives are increasingly being deplatformed from major social media platforms, such as Twitter, Facebook, TikTok and YouTube. Deplatforming has been the subject of extensive debate in terms of whether it unduly constraints free speech as well as whether it is effective in preventing harm or has undesirable negative consequences.[235] Some caution that it is an ineffective 'whack-a-mole' strategy as hateful actors will merely move to sites with less strict rules, taking a portion of their audience with them – and will then be replaced on the more mainstream platforms with other figureheads. But others argue that the policy will decrease the extremity and potency of hate in mainstream spaces and that appetite for cross-platform migration is low.[236] In a study on Reddit, researchers at Georgia Institute of Technology examined how removal of two hateful subreddits (subforums on the site with their own moderators and message boards) impacted the subreddits' users.[237] They found evidence to support both sides of this debate: (1) many users left the site following the bans but (2) that those who stayed engaged in substantially less hate speech. Although this research is limited by the difficulty of measuring online hate accurately (the researchers used keywords to find hateful language, which is a limited method), it indicates that bans may work for each platform but can create wider negative consequences as some users may migrate elsewhere. Deplatforming remains a contentious subject and more studies are needed to fully understand its impact on society, rather than just each platform.

[235] Richard Rogers, "Deplatforming: Following Extreme Internet Celebrities to Telegram and Alternative Social Media", *European Journal of Communication* 35: 3, pp. 213-229 (2020). Available at: https://doi.org/10.1177/0267323120922066.

[236] Hope Not Hate, "Deplatforming works: let's get on with it", 4 October 2019. Available at: https://www.hopenothate.org.uk/2019/10/04/deplatforming-works-lets-get-on-with-it/.

[237] Eshwar Chandrasekharan et al., "You can't stay here: the efficacy of Reddit's 2015 Ban examined through hate speech", *Proceedings of the ACM on Human-Computer Interactions*, 1/2: 31, pp. 1-22 (2017). Available at: https://doi.org/10.1145/3134666

3.6.2 Privacy

Privacy is a key concern online[238] and, following the *Cambridge Analytica* scandal in 2018, has been at the fore of public debates about the regulation of online platforms. Privacy is often associated with anonymity, a closely connected but different concept. Privacy and anonymity can be separated based on (a) whether the person is known and (b) what is known about them: "under the condition of privacy, we have knowledge of a person's identity, but not of an associated personal fact, whereas under the condition of anonymity, we have knowledge of a personal fact, but not of the associated person's identity. In this sense, privacy and anonymity are flip sides of each other." (p. 1755)[239]

Platforms will have different information about their users' demographics, activity and interests. The quality and granularity of this information will depend in part of what the users choose to report to the platforms, their online activities, and the platforms' data practices. Concerns about data use include whether personally identifiable information is used as an input into an AI system (i.e., it is used as a signal in the detection of hate), whether it is stored without the users' explicit permission as part of the moderation process (potentially in 'enriched' form with information about the moderation steps taken), whether it is made visible to individual moderation workers, and whether new forms of personal information are collected solely for the moderation process. Studies show that users highly value their personal data and seek to protect it. The value attributed to this data can vary; data on finances and medical records benefit from greater protections than other forms of data, such as information about individuals' physical activity or energy use.[240] When determining the scope of actions to address

[238] Lemi Baruh et al., "Online Privacy Concerns and Privacy Management: A Meta-Analytical Review", *Journal of Communication*, 67: 1, pp. 26-53 (2017). Available at: https://doi.org/10.1111/jcom.12276.
[239] Jeffrey Skopek, "Anonymity, the Production of Goods and Institutional Design", *Fordham Law Review*, 82: 4, pp. 1751-1809. Available at: https://ir.lawnet.fordham.edu/flr/vol82/iss4/4
[240] Anya Skatova et al., "Unpacking Privacy: Willingness to Pay to Protect Personal Data," *PsyArXiv* (2019). Available at: https://psyarxiv.com/ahwe4/.

hateful content (and the amount and nature of personal data that is monitored, analysed and stored) privacy needs to be at the forefront of considerations.[241]

Ensuring personal data protection is the role of the UK's privacy regulator, the Information Commissioner's Office (ICO).[242] ICO is independent of the UK government and "upholds information rights in the public interest, promoting transparency and accountability by public bodies and organisations and protecting individuals' privacy and information access rights." [243] It is responsible for promoting and enforcing the General Data Protection Regulation (GDPR), the Data Protection Act 2018 (DPA18), the Freedom of Information Act 2000 (FOIA), the Privacy and Electronic Regulations 2003 (PECR) and the Environmental Information Regulations 2004 (EUR). The use of personal data in relation to hateful content falls within ICO's remit.

3.7 Other approaches for tackling online hate

3.7.1 Social psychological theories

A large body of scholarship has researched the socio-psychological and demographic origins of prejudice, such as the role played by contact (or lack thereof) between groups,

[241] See also: The Alan Turing Institute, *Response of the Public Policy Programme to the DCMS and the Home Office's Online Harms White Paper* (London: The Alan Turing Institute, 2019). Available at: https://www.turing.ac.uk/sites/default/files/2019-07/response_of_the_public_policy_programme_to_the_dcms_and_the_home_offices_online_harms_white_paper.pdf.

[242] Ofcom, *Call for Evidence: Video-sharing Platform Regulation* (London: Ofcom, 2020). Available at: https://www.ofcom.org.uk/consultations-and-statements/category-1/video-sharing-platform-regulation.

[243] Elizabeth Denham, *The Information Commissioner's response to the Department for Digital, Culture, Media & Sport consultation on the Online Harms White Paper* (London: The Information Commissioner's Office, 2019). Available at: https://ico.org.uk/media/about-the-ico/consultation-responses/2019/2615232/ico-response-online-harms-20190701.pdf.

as well as the role of economic, social, cultural and political competition.[244] Increasingly, the potential of online contact (or 'e-contact') to reduce prejudice has also been explored.[245] Social psychological research, both online and off-, has mostly focused on the causes of prejudicial *attitudes* (i.e., holding hateful, bigoted or offensive views about a group) rather than *actions*, such as producing hateful content. This is a key distinction and substantial research shows evidence of a gap between attitudes and actions, highlighting the complex pathways which lead individuals from one to the other.[246] An individual may hold prejudicial beliefs but never engage in any hate speech or, alternatively, may hold relatively tolerant beliefs but make a one-off bigoted or derogatory remark about a group. Further, the socio-technical nature of online platforms means that understanding individuals' traits and outlooks will only explain so much, as platform design also plays a key role in how content is produced, shared and engaged with.[247] These theoretical problems are compounded by the difficulty of accessing appropriate

[244] Miles Hewstone et al., "Intergroup Bias", *Annual Review of Psychology*, 53: 1, pp. 575–604 (2002). Available at: https://doi.org/10.1146/annurev.psych.53.100901.135109.
Rupert Brown and Miles Hewstone, "An Integrative Theory of Intergroup Contact", pp. 255-343 in P. Zanna (eds.) *Advances in Experimental Social Psychology,* (San Diego: Elsevier Academic Press, 2005). Available at: https://doi.org/10.1016/S0065-2601(05)37005-5.

[245] Nuri Kim and Magdalena Wojcieszak, "Intergroup contact through online comments: effects of direct and extended contact on outgroup attitudes", *Computers in Human Behaviour*, 81: 1, pp. 63-72 (2018) Available at: https://doi.org/10.1016/j.chb.2017.11.013;
Fiona White, Lauren Harvey and Hisham Abu-Rayya, "Improving intergroup relations in the Internet age: a critical review", *Review of General Psychology*, 19: 2, pp. 129-139 (2015) Available at: https://doi.org/10.1037/gpr0000036.

[246] Clark McCauley and Sophia Moskalenko, "Mechanisms of Political Radicalization: Pathways Toward Terrorism", *Terrorism and Political Violence*, 20: 3, pp. 415-433 (2008). Available at: https://doi.org/10.1080/09546550802073367.
Martin Fishbein and Icek Ajzen*, Predicting and Changing Behavior* (New York: Psychology Press, 2010). Available at: https://doi.org/10.4324/9780203838020.

[247] Carolin Gerlitz and Celia Lury, "Social Media and Self-Evaluating Assemblages: On Numbers, Orderings and Values", *Distinktion: Journal of Social Theory*, 15: 2, pp. 174-188 (2014). Available at: https://doi.org/10.1080/1600910X.2014.920267;
Durkin Mark et al., "A Socio-Technical Perspective on Social Media Adoption: A Case from Retail Banking", *International Journal of Bank Marketing* 33: 7, pp. 944-962 (2015). Available at: https://doi.org/10.1108/IJBM-01-2015-0014.;

data and measuring the concepts used in social psychological theories of prejudice in an online context. Most research is also experimental in design which can introduce certain biases.[248]

3.7.2 Media literacy

Media literacy has been proposed as a potential way of tackling the spread of harmful content online, such as hate speech. [249] Media literacy can be defined and implemented in different ways, according to how expansively/narrowly it is viewed and whether it is being tied to a particular initiative.[250] We define 'media literacy' in line Ofcom's definition, "the ability to use, understand and create media and communications in a variety of contexts" (p. 2).[251] Note that media literacy has been closely linked with digital literacy and information literacy, and in an increasingly digitised world these different literacies often overlap.

Media literacy can be used as a tool to tackle online hate because it has the potential to increase users' resilience and critical faculties. This, in turn, could enable them to counter and challenge not just hate but other harmful content. Researchers at the Institute for Strategic Dialogue argue that "Rather than solely focusing efforts to stop

[248] Oliver Christ and Ulrich Wagner, "Methodological Issues in the study of intergroup contact: towards a new wave of research" in G. Hodson and M. Hewstone (eds.), *Advances in intergroup contact* (Hove: Psychology Press, 2013).

[249] Renee Hobbs, *Digital and Media Literacy: A Plan of Action* (Washington: The Aspen Institute, 2010). Available at: https://kf-site-production.s3.amazonaws.com/publications/pdfs/000/000/075/original/Digital_and_Media_Literacy_A_Plan_of_Action.pdf.

[250] Provision of tools and information for media literacy is mentioned explicitly in the AVMSD in Measure 8 (see Part 1).

[251] David Buckingham, *The Media Literacy of Children and Young People: a review of the research literature on behalf of Ofcom* (London: Ofcom, 2005). Available at: https://discovery.ucl.ac.uk/id/eprint/10000145/;
See also: Ofcom, *Ofcom's Strategy and Priorities for the Promotion of Media Literacy* (London: Ofcom, 2004). Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0021/72255/strat_prior_statement.pdf.

young people coming into contact with these views, we need to give them the critical thinking and media literacy skills to see through them."[252] Further, there is already a well-established infrastructure to draw on when creating and monitoring media literacy initiatives and numerous academic researchers working in this area.[253] For instance, Ofcom's 'Making Sense of Media' project[254] draws attention to the benefits, challenges and opportunities inherent in improving media literacy.

Understanding the efficacy of media literacy in tackling online hate is a difficult task, which is in need of further research. One challenge is that media literacy can take many forms, from providing users with training and information about being online to changing the design and functionality of online platforms. In terms of individual-level interventions, recent research evidence indicates that media literacy could have a positive effect in tackling other online harms, such as misinformation. A study published in 2020 showed that a digital media literacy intervention increased participants' ability to separate mainstream from false news in both America and India.[255] Researchers presented people with tips to help spot false news stories, which helped them to discern between low- and high-quality news. Other research indicates that individuals with lower digital literacy are more likely to believe false health-related content.[256]

[252] Louise Reynolds, "Defeating hate speech online", *Institute for Strategic Dialogue.* Available at: https://www.isdglobal.org/defeating-hate-speech-online/. Last accessed on 4 December 2020.
[253] For example, note the work of Sonia Livingstone at LSE.
Sonia Livingstone profile. Available at: https://www.lse.ac.uk/media-and-communications/people/academic-staff/sonia-livingstone. Last accessed on 4th December 2020.
[254] Ofcom, "Making Sense of Media". Available at: https://www.ofcom.org.uk/research-and-data/media-literacy-research/publications. Last accessed on 4 December 2020.
[255] Andrew Guess et al., "A digital media literacy intervention increases discernment between mainstream and false news in the United States and India", *PNAS*, 117: 27, pp. 15536-15545 (2020). Available at: https://doi.org/10.1073/pnas.1920498117.
[256] Jon Roozenbeek et al., "Susceptibility to misinformation about COVID-19 around the world", *Royal Society Open Science*, 7: 10, pp. 1-15 (2020). Available at: https://royalsocietypublishing.org/doi/10.1098/rsos.201199.

To our knowledge there are no large-scale, quantitative and longitudinal studies which assess whether media literacy (a) improves users' ability to challenge online hate, (b) reduces their hateful behaviour or (c) increases their robustness to the effects of hate. Without this evidence base, assessments of media literacy's efficacy are by nature somewhat speculative and incomplete. A further concern is that most forms of media literacy are only likely to address part of the problem posed by online hate. For instance, individual-level interventions may help targets of hate and users "at risk" of becoming hateful[257] but are likely to do little to address the more committed and entrenched purveyors of hate.

3.7.3 Counterspeech

Counterspeech can be understood as content which challenges, undermines or otherwise criticises and calls out hateful content. It has attracted support from advocates of free speech who view it as a way to contest and challenge hate without needing to constrain freedom of expression. This view can be summarised as: *the best way to tackle bad speech is through more good speech*.[258] Counterspeech has numerous supporters. In a UNESCO report, Gagliardone et al. argue that "Counter-speech is generally preferable to suppression of speech." (p. 5).[259] Similarly, Bartlett and Krasodomski-Jones contend that counterspeech is "faster, more flexible and responsive, capable of dealing with extremism from anywhere and in any language and retains the principle of free and open public spaces for debate."[260] Perhaps unsurprisingly, counterspeech has also

[257] Ian Brown and Josh Cowls, *Check the Web: Assessing the Ethics and Politics of Policing the Internet for Extremist Material* (Oxford: Oxford Internet Institute, 2015). Available at: https://www.voxpol.eu/wp-content/uploads/2015/11/VOX-Pol_Ethics_Politics_PUBLISHED.pdf

[258] Nadine Strossen, *HATE: Why We Should Resist it with Free Speech, Not Censorship* (Oxford: Oxford University Press, 2018);
Jeffrey W Howard, "Terror, Hate and the Demands of Counter-Speech," *British Journal of Political Science*, pp. 1-16 (2019). Available at: https://doi.org/10.1017/S000712341900053X.

[259] Iginio Gagliardone et al., *Countering online hate speech*.

[260] Jamie Bartlett and Alex Krasodomski-Jones, *Counter-speech: Examining content that challenges extremism online* (London: Demos, 2015). Available at: https://www.demos.co.uk/wp-content/uploads/2015/10/Counter-speech.pdf.

received support from international government institutions, such as the Council of Europe[261], numerous civil society organisations, such as the Anti-Defamation League,[262] and the platforms themselves, such as Facebook.[263]

Whether counterspeech is successful in changing the viewpoint of the hateful content purveyor will depend upon many factors, including how the counter-speaker engages with them, their own attributes and the outlook and proclivities of the hateful content purveyor: a committed racist may not be deterred by being told that their content has caused another harm, and could even become more motivated. Yet in many cases this is not the goal of counterspeech. Its purpose is typically not to change the mindset of the hater but to support the victim. As Benesch elaborates, "We often think that counter-speakers are primarily trying to impact the behaviour or the views of the hateful speakers to whom they are responding [...] They are actually trying to do something different. They are trying to reach the larger reading audience or have a positive impact on the discourse within particular online spaces."[264]

As with any intervention, counterspeech entails costs. It is typically created by socially-minded individuals or by community groups, many of whom have either experienced online hate or are actively involved in efforts to tackle it. This means there is a risk that counterspeech places a burden on users who may have *already been burdened* by being targeted by online hate. This has been criticised for compounding the unfairness of online hate: individuals who suffer the harm it inflicts are also made responsible for

[261] Agata de Latour et al., *WE CAN! Taking Action against Hate Speech through Counter and Alternative Narratives* (Hungary: Council of Europe, 2017). Available at: https://www.coe.int/en/web/no-hate-campaign/we-can-alternatives.

[262] The Anti-Defamation League, "Best Practices" and Counterspeech Are Key to Combating Online Harassment", 7 March 2016. Available at:https://www.adl.org/blog/best-practices-and-counterspeech-are-key-to-combating-online-harassment.

[263] Facebook, "Counterspeech". Available at: https://counterspeech.fb.com/en/. Last accessed on 4 December 2020.

[264] Daniel Jones and Susan Benesch, "Combating Hate Speech Through Counterspeech" (Boston: Harvard Berkman Center for Internet & Society, 2019). Available at: https://cyber.harvard.edu/story/2019-08/combating-hate-speech-through-counterspeech.

challenging it. Matsuda et al. comment on a similar issue regarding the need for victims to provide evidence they have been targeted by hate; it is a "psychic tax imposed on those least able to pay." (p. 18)[265] A related concern is that the focus on counterspeech shifts attention away from the structural factors which enable online hate to be posted, shared and to reach large (and potentially vulnerable) audiences, instead emphasising how individuals can address it. Other concerns pertain to the risks inflicted on the counter-speakers themselves. People who respond to purveyors of online hate with counterspeech may also put themselves at risk of being personally targeted.[266]

Finally, it is worth noting the potential of using bots to automatically generate counterspeech[267], especially the latest generation of sophisticated 'chat' bots.[268] Bots are appealing because they could minimise the amount of time that humans are exposed to hateful content and overcome the fact that many forms of counterspeech do not scale well because they involve humans individually reaching out to potential haters and/or challenging their content. However, it also risks other problems such as accidentally pushing some users to become *more* hateful or reducing users' trust in platforms' policies. Bots can easily be poorly designed, learn bad habits or be exploited by adversaries.[269] Notably, Tay, a chatbot from Microsoft which learned from the users it

[265] Mari Matsuda et al. (eds.), *Words that Wound: Critical Race Theory, Assaultive Speech and the First Amendment*.

[266] Alice Marwick et al., *Best Practices for Conducting Risky Research and Protecting Yourself from Online Harassment* (New York: Data & Society Research Institute, 2016). Available at: https://datasociety.net/pubs/res/Best_Practices_for_Conducting_Risky_Research-Oct-2016.pdf.

[267] Kevin Munger, "Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment", *Political Behavior*, 39: 3, pp. 629-649 (2017). Available at: https://doi.org/10.1007/s11109-016-9373-5; Michał Bilewicz et al., "Artificial intelligence against hate: intervention reducing verbal aggression in the social network environment', *Aggressive Behaviour*, pp. 1-7 (2021) Available at: https://doi.org/10.1002/ab.21948.

[268] Tim Adams, "The charge of the chatbots: how do you tell who's human online?", *The Guardian*, 18 November 2018. Available at: https://www.theguardian.com/technology/2018/nov/18/how-can-you-tell-who-is-human-online-chatbots.

[269] Jing Xu et al., "Recipes for Safety in Open-domain Chatbots", *Arxiv:2010.07079v2* (2020). Available at: https://arxiv.org/pdf/2010.07079.pdf.

interacted with, had to be retired after just one day because it was attacked by trolls, quickly learning from them to create offensive, sexist and racist content.[270] In a sensitive domain such as online hate this is especially important, and the use of counterspeech bots requires far more ethical and social consideration.

[270] Marie Wolf et al., “Why we should have seen this coming: Comments on Microsoft’s Tay ‘Experiment’ and Wider Implications”, *The ORBIT Journal*, 1: 2, pp. 1-12 (2012). Available at: https://doi.org/10.29297/orbit.v1i2.49

Appendices

Appendix A: Definitions of online hate from platforms (December 2020)

These platforms are included because they were identified as popular platforms in the UK in 2020 (p. 109).[271] Their inclusion is not a reflection on whether or not they are likely to be affected by AVMSD regulation.

Facebook (and Instagram)	“We define hate speech as a direct attack on people based on what we call protected characteristics — race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability. We protect against attacks on the basis of age when age is paired with another protected characteristic, and also provide certain protections for immigration status. We define attack as violent or dehumanising speech, harmful stereotypes, statements of inferiority, or calls for exclusion or segregation. We separate attacks into three tiers of severity."[272]
Twitter	“You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories.” [273]

[271] Ofcom, “Online Nation 2020 Report”.

[272] Facebook, “Hate Speech”. Available at: https://www.facebook.com/communitystandards/hate_speech. Last accessed on 4 December 2020.

[273] Twitter, “Hateful Conduct Policy”. Available at: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy. Last accessed on 4 December 2020.

Twitch	“Hateful conduct is any content or activity that promotes, encourages, or facilitates discrimination, denigration, objectification, harassment, or violence based on the following characteristics, and is strictly prohibited: race, ethnicity, or national origin; religion; sex, gender, or gender Identity; sexual orientation; age; disability or serious medical condition; veteran Status”[274]
Vimeo	"We do not allow hateful and discriminatory speech. We define this as any expression that (1) is directed to an individual or group of individuals based upon personal characteristics of that individual or group; (2) conveys a message of inferiority or contempt; and (3) would be considered extremely offensive to a reasonable person. Personal characteristics are core elements of identity that are shared by groups of people (and are generally not specific to any one person) and include: Race, Color, National Origin, and Ethnicity, Gender identity, Sexual Orientation, Religion, Disability, Age."[275]
Imgur	“Attacks on people based on their race, ethnicity, national origin, religion, sex, gender, sexual orientation, age, disability or medical condition; Glorification or endorsement of hateful content or ideologies.”[276]

[274] Twitch, “Hateful Conduct and Harassment”. Available at: https://www.twitch.tv/p/en-gb/legal/community-guidelines/harassment/#hateful-conduct. Last accessed on 4 December 2020.

[275] Vimeo, “Vimeo Acceptable Use Community Guidelines”. Available at: https://vimeo.com/help/guidelines. Last accessed on 4 December 2020.

[276] Imgur, “Abuse, Hate Speech and Harassment”. Available at: https://help.imgur.com/hc/en-us/articles/360029650371-Abuse-Hate-Speech-and-Harassment. Last accessed on 4 December 2020.

LiveLeak	“We do not allow hate speech and bigotry and will remove content promoting violence or hatred against individuals or groups.”[277]
TikTok	“We define hate speech as content that does or intends to attack, threaten, incite violence against, or dehumanise an individual or a group of individuals on the basis of protected attributes. We also do not allow content that verbally or physically threatens violence or depicts harm to an individual or a group based on any of the following protected attributes: Race, Ethnicity National origin Religion Caste Sexual orientation Sex Gender Gender identity Serious disease or disability Immigration status.”[278]
Snapchat	“Hate speech or content that demeans, defames or promotes discrimination or violence on the basis of race, colour, caste, ethnicity, national origin, religion, sexual orientation, gender identity, disability, or veteran status, immigration status, socio-economic status, age, weight or pregnancy status is prohibited.”[279]
YouTube	“We remove content promoting violence or hatred against individuals or groups based on any of the following attributes: Age, Caste, Disability, Ethnicity, Gender Identity and Expression, Nationality, Race, Immigration

[277] LikeLeak, “Liveleak Content and Comment Rules”. Available at: https://www.liveleak.com/rules. Last accessed on 4 December 2020.

[278]TikTok, “Community Guidelines”. Available at: https://www.tiktok.com/community-guidelines?lang=en. Last accessed on 4 December 2020.

[279] Snapchat, “Community Guidelines”. Available at: https://www.snap.com/en-GB/community-guidelines. Last accessed on 4 December 2020.

	Status, Religion, Sex/Gender, Sexual Orientation, Victims of a major violent event and their kin, Veteran status"[280]
Tumblr	"Don't encourage violence or hatred. Don't post content for the purpose of promoting or inciting the hatred of, or dehumanising, individuals or groups based on race, ethnic or national origin, religion, gender, gender identity, age, veteran status, sexual orientation, disability or disease. If you encounter content that violates our hate speech policies, please report it."[281]
Reddit	"Remember the human. Reddit is a place for creating community and belonging, not for attacking marginalised or vulnerable groups of people. Everyone has a right to use Reddit free of harassment, bullying, and threats of violence. Communities and people that incite violence or that promote hate based on identity or vulnerability will be banned. Marginalised or vulnerable groups include, but are not limited to, groups based on their actual and perceived race, color, religion, national origin, ethnicity, immigration status, gender, gender identity, sexual orientation, pregnancy, or disability. These include victims of a major violent event and their families."[282]

Table 1: Definitions of online hate from platforms

[280] YouTube, "Hate Speech Policy". Available at: https://support.google.com/youtube/answer/2801939?hl=en-GB. Last accessed on 4 December 2020.

[281] Tumblr, "Community Guidelines". Available at: https://www.tumblr.com/policy/en/community. Last accessed on 4 December 2020.

[282] Reddit, "Promoting Hate Based on Identity or Vulnerability". Available at: https://www.reddithelp.com/hc/en-us/articles/360045715951. Last accessed on 4 December 2020.

Appendix B: AVMSD measures

a) including and applying in the terms and conditions of the video-sharing platform services the requirements referred to in paragraph 1;

b) including and applying in the terms and conditions of the video-sharing platform services the requirements set out in Article 9(1) for audiovisual commercial communications that are not marketed, sold or arranged by the video-sharing platform providers;

c) having a functionality for users who upload user-generated videos to declare whether such videos contain audiovisual commercial communications as far as they know or can be reasonably expected to know;

d) establishing and operating transparent and user-friendly mechanisms for users of a video-sharing platform to report or flag to the video-sharing platform provider concerned the content referred to in paragraph 1 provided on its platform;

e) establishing and operating systems through which video-sharing platform providers explain to users of video-sharing platforms what effect has been given to the reporting and flagging referred to in point (d);

f) establishing and operating age verification systems for users of video-sharing platforms with respect to content which may impair the physical, mental or moral development of minors;

g) establishing and operating easy-to-use systems allowing users of video-sharing platforms to rate the content referred to in paragraph 1;

h)	providing for parental control systems that are under the control of the end-user with respect to content which may impair the physical, mental or moral development of minors;
i)	establishing and operating transparent, easy-to-use and effective procedures for the handling and resolution of users' complaints to the video-sharing platform provider in relation to the implementation of the measures referred to in points (d) to (h);
j)	providing for effective media literacy measures and tools and raising users' awareness of those measures and tools.

Table 2: Summary of measures in the revised 2018 AVMSD, European legislation[283]

[283] EU Legislation, *Directive (EU) 2018/1808 of the European Parliament and of the Council of 14 November 2018 amending Directive 2010/13/EU*.

References

Adams, Tim. “The charge of the chatbots: how do you tell who’s human online?”, *The Guardian*, 18 November 2018. Available at: https://www.theguardian.com/technology/2018/nov/18/how-can-you-tell-who-is-human-online-chatbots.

Alexander, Julia "Youtube Will Let Steven Crowder Run Ads After Year-Long Suspension For Harassment", *The Verge*, 12 August 2020. Available at: https://www.theverge.com/2020/8/12/21365601/youtube-steven-crowder-monetization-reinstated-harassment-carlos-maza.

Alexander, Julia. “YouTube revokes ads from Steven Crowder until he stops linking to his homophobic T-shirts”, *The Verge,* 5 June 2019. Available at: https://www.theverge.com/2019/6/5/18654196/steven-crowder-demonetized-carlos-maza-youtube-homophobic-language-ads.

Allan, Richard. “Hard Questions: Who Should Decide What Is Hate Speech in an Online Global Community?”, *Facebook* (San Francisco: Facebook, 27 June 2017). Available at: https://about.fb.com/news/2017/06/hard-questions-hate-speech/.

Allport, Gordon. *The nature of prejudice* (New York: Addison-Wesley, 1954).

AllSides. “Don’t be fooled by media bias and fake news.” (2020). Available at: https://www.allsides.com/unbiased-balanced-news. Last accessed on 4 December 2020.

Argawal, Swati and Ashish Sureka. “A focused crawler for mining hate and extremism promoting videos on YouTube”*, Proceedings of the 25th ACM on Hypertext and Social Media*, pp. 294-296 (2014). Available at: https://doi.org/10.1145/2631775.2631776

Article 19. “#MissingVoices”. Available at: https://www.article19.org/campaigns/missingvoices/. Last accessed on 4 December 2020.

Article 19. *Prohibiting incitement to discrimination, hostility or violence* (London: Article19, 2012). Available at: https://www.article19.org/data/files/medialibrary/3548/ARTICLE-19-policy-on-prohibition-to-incitement.pdf.

Assimakopoulos, Stavros et al. “Young People’s Perception of Hate Speech”, pp. 53-85 in Stavros Assimaopoulos et al. (eds.), Online Hate Speech in the European Union: *A Discourse-Analytic Perspective* (Berlin: Springer, 2017). Available at: https://link.springer.com/chapter/10.1007/978-3-319-72604-5_4.

Australia Legislation. *Criminal Code Amendment (Sharing of Abhorrent Violent Material) Bill 2019* (Canberra: Australia, 2019). Available at: https://www.legislation.gov.au/Details/C2019A00038.

Awan, Imran and Irene Zempi. "‘I Will Blow Your Face Off’—Virtual And Physical World Anti-Muslim Hate Crime", *British Journal Of Criminology,* 57: 2, pp. 362-380 (2017). Available at: https://doi.org/10.1093/bjc/azv122.

Awan, Imran and Irene Zempi. "The Affinity Between Online And Offline Anti-Muslim Hate Crime: Dynamics And Impacts", *Aggression And Violent Behavior* 27: 1, pp. 1-18 (2016). Available at: https://www.sciencedirect.com/science/article/abs/pii/S1359178916300015.
Awan, Imran. “Islamophobia and Twitter: a typology of online hate against Muslims on social media”, *Policy & Internet*, 6: 2, pp. 133-150 (2014). Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/1944-2866.POI364.

Babvey, Pouria et al., “Using Social Media Data for Assessing Children’s Exposure to Violence during the COVID-19 Pandemic”, *Child Abuse & Neglect* [in proof], (2020). Available at: https://doi.org/10.1016/j.chiabu.2020.104747.

Bacon, Alison et al. “Understanding public attitudes to hate: developing and testing a U.K. version of the hate crime beliefs scale”, *Journal of Interpersonal Violence*, 0:0, pp. 1-26 (2020) Available at: https://doi.org/10.1177/0886260520906188.

Balica, Raluca. “The Criminalisation of Online Hate Speech: It’s Complicated”, *Contemporary Readings in Law and Social Justice*, 2:1, pp. 184-190 (2017). Available at: https://www.ceeol.com/search/article-detail?id=589426*.*

Barendt, Eric. “What is the Harm of Hate Speech?”, *Ethical Theory and Moral Practice*, 22:3, pp. 539-553 (2019). Available at: https://doi.org/10.1007/s10677-019-10002-0

Barnidge, Matthew et al. “Perceived exposure to and avoidance of hate speech in various communication settings”, *Telematics and Informatics*, 44 (2019). Available at: https://www.sciencedirect.com/science/article/abs/pii/S0736585319307555.

Barrett, Paul M. *Who Moderates the Social Media Giants? A Call to End Outsourcing* (New York: NYU Stern Center for Business and Human Rights, 2020). Available at: https://issuu.com/nyusterncenterforbusinessandhumanri/docs/nyu_content_moderation_report_final_version

Bartlett, Jamie and Alex Krasodomski-Jones. *Counter-speech: Examining content that challenges extremism online* (London: Demos, 2015). Available at: https://www.demos.co.uk/wp-content/uploads/2015/10/Counter-speech.pdf.

Bartlett, Jamie et al. *Anti-social media* (London: Demos, 2014). Available at: https://www.demos.co.uk/files/DEMOS_Anti-social_Media.pdf

Baruh, Lemi et al. “Online Privacy Concerns and Privacy Management: A Meta-Analytical Review”, *Journal of Communication*, 67: 1, pp. 26-53 (2017). Available at: https://doi.org/10.1111/jcom.12276.

BBC News, "Tom Daley 'Abuse' Tweet: Legal Rethink On Online Rules", 20 September 2012. Available at: https://www.bbc.co.uk/news/uk-19660415.

BBC News, "Tom Daley Tweet: No Action Against Daniel Thomas", 20 September 2012. Available at: https://www.bbc.co.uk/news/uk-wales-19661950.

BBC News, “Aristocrat guilty over 'menacing' Gina Miller Facebook post”, July 11 2017. Available at: https://www.bbc.co.uk/news/uk-40574754

BBC News, “Capitol riots: How a Trump rally turned deadly”, 7 January 2020. Available at: https://www.bbc.co.uk/news/av/world-us-canada-55569495.

BBC News, “UK unveils extremism blocking tool”, 13 February 2018. Available at: https://www.bbc.co.uk/news/technology-43037899.

BBC, “Own It, The App: Six Technical Challenges”, 18 September 2019. Available at: https://www.bbc.co.uk/blogs/internet/entries/94ec41ae-b25b-4e58-9c0f-1b9b2890c281.

Benesch, Susan*. What is Dangerous Speech?* (Washington: Dangerous Speech, 2020). Available at: https://dangerousspeech.org/about-dangerous-speech/.

Bhargava, Vikram and Manuel Velasquez. “Ethics of the Attention Economy: The Problem of Social Media Addiction”, *Business Ethics Quarterly*, pp. 1-39 (2020). Accessed at: https://doi.org/10.1017/beq.2020.32

Bhat, Prashanth and Ofra Klein. “Covert Hate Speech: White Nationalists and Dog Whistle Communication on Twitter”, pp.151-172 in Gwen Bouvier and Judith Rosenbaum (eds.) *Twitter, The Public Sphere and the Chaos of Online Deliberation* (Switzerland: Palgrave MacMillan, 2020).

Bickert, Monika. “Removing Holocaust Denial Content” (San Francisco: Facebook, 12 October 2020). Available at: https://about.fb.com/news/2020/10/removing-holocaust-denial-content/.

Big Brother Watch. *Big Brother Watch’s response to the Online Harms White Paper Consultation* (London: Big Brother Watch, 2019). Available at: https://bigbrotherwatch.org.uk/wp-content/uploads/2020/02/Big-Brother-Watch-consultation-response-on-The-Online-Harms-White-Paper-July-2019.pdf.

Bilewicz, Michał et al. “Artificial intelligence against hate: intervention reducing verbal aggression in the social network environment’, *Aggressive Behaviour*, pp. 1-7 (2021) Available at: https://doi.org/10.1002/ab.21948.

BitChute. “Community Guidelines: Content Sensitivity”. Available at: https://support.bitchute.com/policy/guidelines/#content-sensitivity. Last accessed on 4 December 2020.

Bolinger, Renée Jorgensen. “The Pragmatics of Slurs”, *Nous*, 51: 3, pp. 439-462 (2017). Available at: https://doi.org/10.1111/nous.12090

Breeden, Aurelien. “French Court Strikes Down Most of Online Hate Speech Law”, *New York Times*, 18 June 2020, https://www.nytimes.com/2020/06/18/world/europe/france-internet-hate-speech-regulation.html.

British Board of Classification. “Half of children and teens exposed to harmful online content while in lockdown”, 4 May 2020. Available at: https://www.bbfc.co.uk/about-us/news/half-of-children-and-teens-exposed-to-harmful-online-content-while-in-lockdown.

Brown, Alexander and Adriana Sinclair. *The Politics of Hate Speech Laws* (London: Routledge, 2019).

Brown, Alexander. "What Is So Special About Online (As Compared to Offline) Hate Speech?", *Ethnicities*, 18: 13, pp. 297-326 (2017). Available at: https://doi.org/10.1177/1468796817709846

Brown, Alexander. *Models of Governance of Online Hate Speech* (Brussels: Council of Europe, 2020). Available at: https://rm.coe.int/models-of-governance-of-online-hate-speech/16809e671d.

Brown, Ian and Josh Cowls. *Check the Web: Assessing the Ethics and Politics of Policing the Internet for Extremist Material* (Oxford: Oxford Internet Institute, 2015).

Available at: https://www.voxpol.eu/wp-content/uploads/2015/11/VOX-Pol_Ethics_Politics_PUBLISHED.pdf

Brown, Rupert and Miles Hewstone, "An Integrative Theory of Intergroup Contact", pp. 255-343 in P. Zanna (eds.) *Advances in Experimental Social Psychology,* (San Diego: Elsevier Academic Press, 2005). Available at: https://doi.org/10.1016/S0065-2601(05)37005-5.

Brustein, Joshua. "Facebook Grappling With Employee Anger Over Moderator Conditions", *Bloomberg*, 25 February 2019. Available at: https://www.bloomberg.com/news/articles/2019-02-25/facebook-grappling-with-employee-anger-over-moderator-conditions.

Buckingham, David. *The Media Literacy of Children and Young People: a review of the research literature on behalf of Ofcom* (London: Ofcom, 2005). Available at: https://discovery.ucl.ac.uk/id/eprint/10000145/.

Buolamwini, Joy and Timnit Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification", *Proceedings of Machine Learning Research*, 81, pp. 1-15 (2018). Available at: http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

Burden, Emily. Profile on Southampton University's website. Available at: https://www.southampton.ac.uk/history/postgraduate/research_students/elb1g13.page. Last accessed on 15 February 2021.

Butler, Judith. *Excitable Speech: A Politics of the Performative* (London: Routledge, 1997).

Cambridge Consultants. *Use of AI in online content moderation* (London: Ofcom, 2019). Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf.

Cave, Damien. “Australia Passes Law to Punish Social Media Companies for Violent Posts”, *New York Times,* 3 April 2019. Available at: https://www.nytimes.com/2019/04/03/world/australia/social-media-law.html.
Cecez-Kecmanovic, Dubravka. “The sociomateriality of information systems: current status, future directions”, *MIS Quarterly,* 38: 3, pp. 809-830. Available at: https://doi.org/10.25300/MISQ/2014/38:3.3.

Cepollaro, Bianca and Dan Zeman. “The challenge from non-derogatory uses of slurs”, *BRILL*, 97: 1, pp.1-10 (2020). Available at: https://doi.org/10.1163/18756735-09701002.

Chandrasekharan, Eshwar et al. “You can’t stay here: the efficacy of Reddit’s 2015 Ban examined through hate speech”, *Proceedings of the ACM on Human-Computer Interactions*, 1/2: 31, pp. 1-22 (2017). Available at: https://doi.org/10.1145/3134666

Chetty, Naganna and Sreejith Alathur. “Hate speech review in the context of online social networks”, *Aggression and Violent Behaviour*, 40, pp. 108-118 (2018) Available at: https://doi.org/10.1016/j.avb.2018.05.003.
Christ, Oliver and Ulrich Wagner. “Methodological Issues in the study of intergroup contact: towards a new wave of research” in G. Hodson and M. Hewstone (eds.), *Advances in intergroup contact* (Hove: Psychology Press, 2013).

Citron, Danielle and Helen Norton. “Intermediaries and hate speech: fostering digital citizenship for our information age”, *Boston University Law Review*, 91:16, pp. 1435-1484 (2011). Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1764004.

Commission for Countering Extremism, *Challenging Hateful Extremism* (London: UK Home Office, 2019). Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/874101/200320_Challenging_Hateful_Extremism.pdf.

Conway, Maura et al. “Down the (White) Rabbit hole: the extreme right and online recommender systems”, *Social Science Computer Review*, 33:4, pp. 459-478 (2015). Available at: https://journals.sagepub.com/doi/pdf/10.1177/0894439314555329.

Costello, Matthew et al. "Social Group Identity And Perceptions Of Online Hate", *Sociological Inquiry*, 89: 3, pp. 427-452 (2019). Available at: https://doi.org/10.1111/soin.12274.

Council of the European Union Press Release. “Joint statement of EU Ministers for Justice and Home Affairs and representatives of EU institutions on the terrorist attacks in Brussels on 22 March 2016”, 24 March 2016. Available at: https://www.consilium.europa.eu/en/press/press-releases/2016/03/24/statement-on-terrorist-attacks-in-brussels-on-22-march/.

Crown Prosecution Service. *Hate Crime Report 2018-2019* (London: Crown Prosecution Service, 2019). Available at: https://www.cps.gov.uk/sites/default/files/documents/publications/CPS-Hate-Crime-Annual-Report-2018-2019.PDF.

Crown Prosecution Service. *Social Media - Guidelines on prosecuting cases involving communications sent via social media* (London: Crown Prosecution Service, 2018). Available at: https://www.cps.gov.uk/legal-guidance/social-media-guidelines-prosecuting-cases-involving-communications-sent-social-media.

Culliford, Elizabeth. “From hate speech to nudity, Facebook's oversight board picks its first cases”, *Reuters*, 1 December 2020. Available at: https://uk.reuters.com/article/facebook-oversight/from-hate-speech-to-nudity-facebooks-oversight-board-picks-its-first-cases-idINKBN28B50Y.

Davidson, Thomas et al. “Racial Bias in Hate Speech and Abusive Language Detection Datasets”, in *Proceedings of the Third Workshop on Abusive Language Online* (Florence: Association for Computational Linguistics), pp. 25-35 (2019). Available at: https://www.aclweb.org/anthology/W19-3504.

De Latour, Agata et al. *WE CAN! Taking Action against Hate Speech through Counter and Alternative Narratives* (Hungary: Council of Europe, 2017). Available at: https://www.coe.int/en/web/no-hate-campaign/we-can-alternatives.

De Streel, Alexandre et al. *Online Platforms' Moderation of Illegal Content Online: Law, Practices and Options for Reform* (Luxembourg: Policy Department for Economic, Scientific and Quality of Life Policies, 2020). Available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/652718/IPOL_STU(2020)652718_EN.pdf.

Denham, Elizabeth. *The Information Commissioner's response to the Department for Digital, Culture, Media & Sport consultation on the Online Harms White Paper* (London: The Information Commissioner's Office, 2019). Available at: https://ico.org.uk/media/about-the-ico/consultation-responses/2019/2615232/ico-response-online-harms-20190701.pdf.

Department for Digital, Culture, Media & Sport and the Home Office. *Online Harms White Paper* (London: UK Government, 2019). Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf.

Department for Digital, Culture, Media & Sport and the Home Office, *Online Harms White Paper: Full Government Response to the consultation*.(London: UK Government, 2020). Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/944310/Online_Harms_White_Paper_Full_Government_Response_to_the_consultation_CP_354_CCS001_CCS1220695430-001__V2.pdf.

Downs, Daniel et al. "Predicting the Importance of Freedom of Speech and the Perceived Harm of Hate Speech", *Journal of Applied Social Psychology*, 42:6, pp. 1353–1375 (2012) Available at: https://doi.org/10.1111/j.1559-1816.2012.00902.x

Dynabench. "Dynabench". Available at: https://dynabench.org/about. Last accessed on 4 December 2020.

Echikson, William and Olivia Knodt, *Germany's NetzDG: A key test for combating online hate* (Brussels: Counter-Extremism Project, 2018). Available at: http://wp.ceps.eu/wp-content/uploads/2018/11/RR%20No2018-09_Germany's%20NetzDG.pdf.

Elmer, Greg. “Prospecting Facebook: the limits of the economy of attention”, *Media, Culture & Society*, 41: 3, pp. 332-346, (2018). Accessed at: https://doi.org/10.1177/0163443718813467

Ernst, Julian et al. “Hate beneath the counter speech?”, *Journal for Deradicalization*, 10: 1, pp. 1-49 (2017). Accessed at: https://journals.sfu.ca/jd/index.php/jd/article/view/91

Eschmann, Rob. “Digital Resistance: How online communication facilitates responses to racial microaggressions”, *Sociology of Race and Ethnicity*, pp. 1-14 (2020). Available at: https://doi.org/10.1177/2332649220933307

EU Legislation. *Television broadcasting activities: "Television without Frontiers" (TVWF) Directive* (Brussels: European Union, 1989). Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=LEGISSUM%3Al24101.

EU Legislation. *Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law* (Brussels: European Union, 2008). Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2008.328.01.0055.01.ENG&toc=OJ:L:2008:328:TOC.

EU Legislation, *Proposal for a Directive of the European Parliament and of the Council amending Directive 2010/13/EU* (Brussels: European Union, 2016). Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1464618463840&uri=COM:2016:287:FIN.

EU Legislation. *Directive 2010/13/EU of the European Parliament and of the Council of 10 March 2010* (Brussels: European Union, 2010). Available at: http://data.europa.eu/eli/dir/2010/13/oj.

EU Legislation. *Directive (EU) 2018/1808 of the European Parliament and of the Council of 14 November 2018 amending Directive 2010/13/EU* (Brussels: European Union, 2018). Available at: http://data.europa.eu/eli/dir/2018/1808/oj.

EU Legislation. *EU Charter of Fundamental Rights 2012* (Brussels: European Union, 2012). Available at: https://fra.europa.eu/en/eu-charter/article/21-non-discrimination.

European Commission against Racism and Intolerance (ECRI). *ECRI General Policy Recommendation No. 15 on Combating Hate Speech* (Strasbourg: Council of Europe, 2016). Available at; https://rm.coe.int/ecri-general-policy-recommendation-no-15-on-combating-hate-speech/16808b5b01

European Commission Press Release. "A Digital Single Market for Europe: Commission sets out 16 initiatives to make it happen", 6 May 2015, https://ec.europa.eu/commission/presscorner/detail/en/IP_15_4919.

European Commission. "Public consultation on Directive 2010/13/EU on Audiovisual Media Services (AVMSD) - A media framework for the 21st century", 6 July 2015 to 30 September 2015. Available at: https://ec.europa.eu/digital-single-market/en/news/public-consultation-directive-201013eu-audiovisual-media-services-avmsd-media-framework-21st.

European Court of Human Rights. *Guide on Article 10 of the European Convention on Human Rights: Freedom of Expression* (Brussels: Council of Europe, 2020). Available at: https://www.echr.coe.int/Documents/Guide_Art_10_ENG.pdf.

European Parliament. "Audiovisual and media policy". Available at: https://www.europarl.europa.eu/factsheets/en/sheet/138/audiovisual-and-media-policy. Last accessed on 4 December 2020.

Facebook Oversight Board. "Announcing the Oversight Board's first cases and appointment of trustees", December 2020. Available at: https://www.oversightboard.com/news/719406882003532-announcing-the-oversight-board-s-first-cases-and-appointment-of-trustees/.

Facebook, "Community Standards: Hate Speech". Available at: https://www.facebook.com/communitystandards/recentupdates/hate_speech/. Last accessed on 4 December 2020.

Facebook, “Counterspeech”. Available at: https://counterspeech.fb.com/en/. Last accessed on 4 December 2020.

Facebook, “Hate Speech”. Available at: https://www.facebook.com/communitystandards/hate_speech. Last accessed on 4 December 2020.

Facebook. “What ‘The Social Dilemma’ gets wrong”, (San Francisco: Facebook, 2020). Available at: https://about.fb.com/wp-content/uploads/2020/10/What-The-Social-Dilemma-Gets-Wrong.pdf.

Facebook. *Community Standards Enforcement Report Q3 2020* (San Francisco: Facebook, 2020). Available at: https://transparency.facebook.com/community-standards-enforcement#hate-speech

Farrell, Tracie et al. “Exploring Misogyny across the Manosphere in Reddit”, in *Proceedings of the 10th ACM Conference on Web Science* (New York: Association for Computing Machinery), pp. 87–96 (2019). Available at: https://doi.org/10.1145/3292522.3326045.

Fishbein, Martin and Icek Ajzen*, Predicting and Changing Behavior* (New York: Psychology Press, 2010). Available at: https://doi.org/10.4324/9780203838020.

French Constitutional Council Press Release. “Decision 2020-801 DC of June 18, 2020 press release”, (Paris: French Constitutional Council, 18 June 2020). Available at: https://www.conseil-constitutionnel.fr/actualites/communique/decision-n-2020-801-dc-du-18-juin-2020-communique-de-presse.

Gagliardone, Iginio et al. *Countering online hate speech* (Paris: UNESCO, 2017). Available at: https://unesdoc.unesco.org/ark:/48223/pf0000233231.

Gallacher, John. “Automated Detection of Terrorist and Extremist Content” in Bharath Ganesh and Jonathan Bright (eds.) *Extreme Digital Speech* (London: Vox-POL, 2019). Available at: https://www.voxpol.eu/download/vox-pol_publication/DCUJ770-VOX-Extreme-Digital-Speech.pdf

Gallagher, Ryan J. et al. “Divergent discourse between protests and counter-protests: #BlackLivesMatter and #AllLivesMatter”, *Plos ONE*, 13: 4, pp. 1-23 (2018). Available at: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0195644.

Gallie, W.B. "Essentially Contested Concepts," *Proceedings of the Aristotelian Society* 56:1, pp. 167-98 (1955). Available at: https://www.jstor.org/stable/4544562.

Ganesh, Bharath. “The ungovernability of digital hate culture”, *Journal of International Affairs*, 72: 2, pp. 30-49 (2018). Available at: https://www.jstor.org/stable/26552328.

George, Cherian. “Hate Speech Law and Policy” in P.H. Ang and R. Mansell (eds.) *The International Encyclopedia of Digital Communication and Society* (New Jersey: Wiley-Blackwell, 2015). Available at: https://doi.org/10.1002/9781118767771.wbiedcs139.

Gerlitz, Carolin and Celia Lury. “Social Media and Self-Evaluating Assemblages: On Numbers, Orderings and Values”, *Distinktion: Journal of Social Theory*, 15: 2, pp. 174-188 (2014). Available at: https://doi.org/10.1080/1600910X.2014.920267.

German Federal Ministry of Justice and Consumer Protection. *Act To Improve Enforcement Of The Law In Social Networks (Network Enforcement Act)* (Berlin: Germany, 2017). Available at: https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.html.

Gillespie, Tarleton et al. “Expanding the Debate about Content Moderation: Scholarly Research Agendas for the Coming Policy Debates.” *Internet Policy Review*, 9: 4 (2020). Available at: https://doi.org/10.14763/2020.4.1512

Gillespie, Tarleton. “Content moderation, AI and the question of scale”, *Big Data & Society*, July-December, pp. 1-5, (2020). Available at: https://doi.org/10.1177/2053951720943234

Gillespie, Tarleton. “Looking beyond Facebook: moderation everywhere” in Tarleton Gillespie et al. “Expanding the Debate about Content Moderation: Scholarly Research

Agendas for the Coming Policy Debates", Internet Policy Review, 9: 4, pp. 4-7 (2020). Available at: https://doi.org/10.14763/2020.4.1512

Glitch!. https://fixtheglitch.org/online-abuse/. Last accessed on 4 December 2020.

Goel, Sharad et al. "The Structural Virality of Online Diffusion", *Management Science*, 62: 1 (2015). Available at: https://doi.org/10.1287/mnsc.2015.2158

Google. "Appeals". Available at: https://transparencyreport.google.com/youtube-policy/appeals?hl=en_GB. Last accessed on 4 December 2020.

Griffin, Nick. Twitter profile, available at: https://twitter.com/NickGriffinBU. Last accessed on 31 December 2020.

Gröndahl, Tommi et al. "All You Need is "Love": Evading Hate Speech Detection", in *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security* (New York: Association for Computing Machinery), pp. 2–12 (2018). Available at: https://doi.org/10.1145/3270101.3270103.

Guess, Andrew et al. "A digital media literacy intervention increases discernment between mainstream and false news in the United States and India", *PNAS*, 117: 27, pp. 15536-15545 (2020). Available at: https://doi.org/10.1073/pnas.1920498117.

Guhl, Jakob and Jacob Davey, "Hosting the 'Holohoax': A Snapshot of Holocaust Denial Across Social Media" (London: Institute for Strategic Dialogue, 2020). Available at: https://www.isdglobal.org/wp-content/uploads/2020/08/Hosting-the-Holohoax.pdf.

Halprin, Matt. "An update to our harassment policy, Official YouTube Blog", *YouTube*, 11 December 2019. Available at: https://blog.youtube/news-and-events/an-update-to-our-harassment-policy/.

Harris, Brent. "Oversight Board to Start Hearing Cases", *Facebook*, 22 October 2020. Available at: https://about.fb.com/news/2020/10/oversight-board-to-start-hearing-cases/.

Hatmaker, Taylor. “Facebook’s controversial Oversight Board starts reviewing content moderation cases”, *TechCrunch*, 22 October 2020. Available at: https://techcrunch.com/2020/10/22/facebook-oversight-board-controversy/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAMWkyH86ZK2RbsAHdUsZL7LJKb0gz6iTeZwDyj1SC5Ye9HIqKwPdsIyyhds1yYq41fg5_zz8UL5AYtTd6t-512dUPPM-Nq8x1_lcW5wDREqoitVq84amL1GfFzuV8K3pXfJsXLf3LSmrGd6QZ3QcVXYV72h6kXBXSPNxq0PgPsEL.

Henry, Jessica. “Beyond free speech: novel approaches to hate on the Internet in the United States”, I*nformation & Communications Technology Law,* 18: 2, pp. 235-251 (2009). Available at: 10.1080/13600830902808127.

Hern, Alex and Julia Carrie Wong. "Facebook Employees Hold Virtual Walkout Over Mark Zuckerberg's Refusal To Act Against Trump", *The Guardian*, 1 June 2020. Available at: https://www.theguardian.com/technology/2020/jun/01/facebook-workers-rebel-mark-zuckerberg-donald-trump.

Hern, Alex. "Facebook And Instagram Ban Antisemitic Conspiracy Theories And Blackface", *The Guardian*, 12 August 2020. Available at: https://www.theguardian.com/technology/2020/aug/12/facebook-and-instagram-ban-antisemitic-conspiracy-theories-and-blackface.

Hewstone, Miles et al. “Intergroup Bias”, *Annual Review of Psychology*, 53: 1, pp. 575–604 (2002). Available at: https://doi.org/10.1146/annurev.psych.53.100901.135109.

Hickman, Martin. "Chief prosecutor reveals lenient stance after footballer is cleared of abusing Tom Daley", *The Independent*, 20 September 2012. Available at: https://www.independent.co.uk/news/uk/crime/chief-prosecutor-reveals-lenient-stance-after-footballer-cleared-abusing-tom-daley-8160648.html.

Hine, Gabriel Emile et al. “Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and its Effects on the We”, in *Proceedings of the Eleventh International AAAI Conference on Web and Social Media* (Montreal:

Association for the Advancement of Artificial Intelligence (AAAI), pp. 92-101 (2017). Available at: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15670.

Hobbs, Renee. *Digital and Media Literacy: A Plan of Action* (Washington: The Aspen Institute, 2010). Available at: https://kf-site-production.s3.amazonaws.com/publications/pdfs/000/000/075/original/Digital_and_Media_Literacy_A_Plan_of_Action.pdf.

Hollister, Sean. "TikTok will now tell you why it removed your video", *The Verge*, 22 October 2020. Available at: https://www.theverge.com/2020/10/22/21529497/tiktok-content-violation-which-policy-community-guidelines-update.

Home Affairs Committee*. Hate crime: abuse, hate and extremism online - Fourteenth Report of Session 2016–17* (London: UK Government, 2017). Available at: https://publications.parliament.uk/pa/cm201617/cmselect/cmhaff/609/609.pdf.

Home Office. *Hate Crime, England and Wales 2017/18* (London: Home Office, 2018). Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/748598/hate-crime-1718-hosb2018.pdf.

HOPE Not Hate. "Deplatforming works: let's get on with it", 4 October 2019. Available at: https://www.hopenothate.org.uk/2019/10/04/deplatforming-works-lets-get-on-with-it/.

HOPE not Hate. *A Better Web: Regulating to Reduce Far-Right Hate Online* (London: HOPE not Hate, 2020). Available at: https://www.hopenothate.org.uk/wp-content/uploads/2020/11/A-Better-Web-Final-.pdf.

Howard, Jeffrey W. "Terror, Hate and the Demands of Counter-Speech," *British Journal of Political Science*, pp. 1-16 (2019). Available at: https://doi.org/10.1017/S000712341900053X.

Imgur, “Abuse, Hate Speech and Harassment”. Available at: https://help.imgur.com/hc/en-us/articles/360029650371-Abuse-Hate-Speech-and-Harassment. Last accessed on 4 December 2020.

Imhoff, Roland and Julia Recker, "Differentiating Islamophobia: Introducing A New Scale To Measure Islamoprejudice And Secular Islam Critique", *Political Psychology*, 33: 6, pp. 811-824 (2012). Available at: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1467-9221.2012.00911.x.
Index on Censorship. *Index on Censorship submission to Online Harms White Paper consultation* (London: Index on Censorship, 2019). Available at: https://www.indexoncensorship.org/wp-content/uploads/2019/07/Online-Harms-Consultation-Response-Index-on-Censorship.pdf.

Instagram. “Account Disable Policy Changes on Instagram”. 18 July 2019. Available at: https://about.instagram.com/blog/announcements/account-disable-policy-changes-on-instagram. Last accessed on 4 December.

Institute for Strategic Dialogue. *A joint statement on the Online Harms White Paper and the direction of regulation in the UK* (London: Institute for Strategic Dialogue, 2020). Available at: https://www.isdglobal.org/isd-publications/joint-statement-on-the-online-harms-white-paper/.

Jay, Timothy. “Do offensive words harm people?” *Psychology, Public Policy, and Law,* 15: 2, pp. 81–101 (2009). Available at: https://doi.org/10.1037/a0015646

Johnson, Neil. F et al. "Hidden Resilience And Adaptive Dynamics Of The Global Online Hate Ecology", *Nature* 573, pp. 261-265 (2019). Available at: https://www.nature.com/articles/s41586-019-1494-7.

Jones, Daniel and Susan Benesch. “Combating Hate Speech Through Counterspeech” (Boston: Harvard Berkman Center for Internet & Society, 2019). Available at: https://cyber.harvard.edu/story/2019-08/combating-hate-speech-through-counterspeech.

Juneja, Preran et al. “Through the Looking Glass: Study of Transparency in Reddit’s Moderation Practices”, *Proceedings of the ACM on Human-Computer Interaction*, 4: 1 (2019). Available at: https://doi.org/10.1145/3375197

Kaakinen, Markus et al. "Did The Risk Of Exposure To Online Hate Increase After The November 2015 Paris Attacks? A Group Relations Approach", *Computers In Human Behavior,* 78: 1, pp. 90-97 (2018). Available at: https://www.sciencedirect.com/science/article/abs/pii/S0747563217305484#.

Kiela, Douwe et al. “The hateful memes challenge: Detecting hate speech in multimodal memes”, *arXiv:2005.04790* (2020). Available at: https://arxiv.org/abs/2005.04790.

Kienpointner, Manfred. “Impoliteness online, hate speech in online interactions”, *Internet Pragmatics*, 1: 2, pp. 329-351 (2018). Available at: https://www.jbe-platform.com/content/journals/10.1075/ip.00015.kie.

Kim, Nuri and Magdalena Wojcieszak. “Intergroup contact through online comments: effects of direct and extended contact on outgroup attitudes”, *Computers in Human Behaviour*, 81: 1, pp. 63-72 (2018) Available at: https://doi.org/10.1016/j.chb.2017.11.013.

Kosenko, Kami. “The hijacked hashtag: the constitutive features of abortion stigmatization. The #ShoutYourAbortion Twitter Campaign”, *International Journal of Communication*, 13: 1, pp. 1-21 (2019). Available at: https://ijoc.org/index.php/ijoc/article/view/7849.

Law Commission. “Consultation on the reform of the communications offences”, 24 September to 18 December 2020. Available at: https://consult.justice.gov.uk/law-commission/online_comms/.

Law Commission. “Hate Crime Consultation”, https://www.lawcom.gov.uk/project/hate-crime/. Last accessed on 4 December 2020.

Law Commission. *Harmful Online Communications: The Criminal Offences - A Consultation Paper* (London: Law Commission, 2020). Available at: https://s3-eu-west-

2.amazonaws.com/lawcom-prod-storage-11jsxou24uy7q/uploads/2020/09/Online-Communications-Consultation-Paper-FINAL-with-cover.pdf.

Law Commission. *Hate Crime: Consultation Paper Summary* (London: Law Commission, 2020). Available at: https://s3-eu-west-2.amazonaws.com/lawcom-prod-storage-11jsxou24uy7q/uploads/2020/09/Hate-crime-final-summary.pdf.

Left/Right News. “Look Left. Look Right. Think Straight.” (2020). Available at: https://leftright.news. Last accessed on 4 December 2020.

Leslie, David*. Understanding artificial intelligence, ethics and safety* (London: The Alan Turing Institute, 2019). Available at: https://www.turing.ac.uk/research/publications/understanding-artificial-intelligence-ethics-and-safety.

LikeLeak, “Liveleak Content and Comment Rules”. Available at: https://www.liveleak.com/rules. Last accessed on 4 December 2020.

Livingstone, Sonia. Profile on LSE website. Available at: https://www.lse.ac.uk/media-and-communications/people/academic-staff/sonia-livingstone. Last accessed on 4th December 2020.

Magu, Rijul et al. “Detecting the Hate Code on Social Media,” in *Proceedings of the Eleventh International AAAI Conference on Web and Social Media* (California: Association for the Advancement of Artificial Intelligence) (2017). Available at: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15604.

Margetts, Helen et al. *Political Turbulence: How Social Media Shape Collective Action*, (Oxford: Princeton University Press, 2016).
Margetts, Helen. Profile on The Alan Turing Institute website. Available at: https://www.turing.ac.uk/people/programme-directors/helen-margetts. Last accessed on 4 December 2020.

Mariconti, Enrico et al., “You Know What to Do”: Proactive Detection of YouTube Videos Targeted by Coordinated Hate Attacks”, *Proceedings of the ACM on Human-Computer Interaction*, 3: 207 (2019). Available at: https://dl.acm.org/doi/10.1145/3359309.

Mark, Durkin et al. “A Socio-Technical Perspective on Social Media Adoption: A Case from Retail Banking”, *International Journal of Bank Marketing* 33: 7, pp. 944-962 (2015). Available at: https://doi.org/10.1108/IJBM-01-2015-0014.
Marwick, Alice and danah boyd. “I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience”, *New Media & Society*, 13: 1, pp. 114-133 (2012). Available at: https://doi.org/10.1177/1461444810365313.

Marwick, Alice et al. *Best Practices for Conducting Risky Research and Protecting Yourself from Online Harassment* (New York: Data & Society Research Institute, 2016). Available at: https://datasociety.net/pubs/res/Best_Practices_for_Conducting_Risky_Research-Oct-2016.pdf.

Matsuda. Mari et al. (eds.) *Words that Wound: Critical Race Theory, Assaultive Speech and the First Amendment* (New York: Westview Press, 1993).

Mayor of London, “How the Violence Reduction Unit is tackling the root causes of crime”, 17 July 2019. Available at: https://www.london.gov.uk/city-hall-blog/how-violence-reduction-unit-tackling-root-causes-crime.

McCauley, Clark and Sophia Moskalenko. “Mechanisms of Political Radicalization: Pathways Toward Terrorism”, *Terrorism and Political Violence*, 20: 3, pp. 415-433 (2008). Available at: https://doi.org/10.1080/09546550802073367.

Mehrabi, Ninareh et al. “A survey on bias and fairness in machine learning”, *arXiv:1908.09635v2* (2019). Available at: https://arxiv.org/pdf/1908.09635.pdf.

Millar, Robyn et al. *Considering the evidence of the impacts of lockdown on the mental health and wellbeing of children and young people within the context of the individual, the family, and education* (Glasgow: Mental Health Foundation, 2020). Available at:

https://www.mentalhealth.org.uk/sites/default/files/MHF%20Scotland%20Impacts%20of%20Lockdown.pdf.

Müller, Karsten and Carlo Schwarz. “Fanning the flames of hate; Social media and hate crime”, *Journal of the European Economic Association* (2020). Available at: https://academic.oup.com/jeea/advance-article-abstract/doi/10.1093/jeea/jvaa045/5917396?redirectedFrom=fulltext.

Munger, Kevin. “Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment”, *Political Behavior*, 39: 3, pp. 629-649 (2017). Available at: https://doi.org/10.1007/s11109-016-9373-5.

Munn, Luke. “Alt-right pipeline: Individual journeys to extremism online”, *First Monday*, 24: 6 (2019). Available at: https://doi.org/10.5210/fm.v24i6.10108.

Nagle, Angela*. Kill All Normies: Online Culture Wars from 4chan and Tumblr to Trump and the altright* (London: Zero Books, 2017)

Newton, Casey. “YouTube Moderators Are Being Forced To Sign A Statement Acknowledging That The Job Can Give Them PTSD”, *The Verge*, 24 January 2020. Available at: https://www.theverge.com/2020/1/24/21075830/youtube-moderators-ptsd-accenture-statement-lawsuits-mental-health.

Newton, Casey. "The Secret Lives Of Facebook Moderators In America", *The Verge*, 25 February 2019. Available at: https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona.

Newton, Casey. “Facebook’s new Oversight Board is a wild new experiment in platform governance”, *The Verge*, 23 October 2020. Available at: https://www.theverge.com/2020/10/23/21530524/facebooks-new-oversight-board-platform-governance.

O’Dea, Conor and Donald Saucier. “Perceptions of racial slurs used by black individuals towards white individuals: derogation or affiliation?”, *Journal of Language and Social Psychology*, 39:5-6, pp. 678-700 (2020). Available at: https://doi.org/10.1177/0261927X2090498

O'Regan, Catherin. "Hate Speech Online: an (intractable) contemporary challenge?", *Current Legal Problems*, 71: 1, pp. 403-429 (2018) Available at: https://doi.org/10.1093/clp/cuy012.

Ofcom. "Making Sense of Media". Available at: https://www.ofcom.org.uk/research-and-data/media-literacy-research/publications. Last accessed on 4 December 2020.

Ofcom. "Ofcom to regulate harmful content online", 15 December 2020. Available at: https://www.ofcom.org.uk/about-ofcom/latest/features-and-news/ofcom-to-regulate-harmful-content-online.

Ofcom. "Online Nation 2020 Report" (London: Ofcom, 2020). Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0027/196407/online-nation-2020-report.pdf.

Ofcom. *Call for Evidence: Video-sharing Platform Regulation* (London: Ofcom, 2020). Available at: https://www.ofcom.org.uk/consultations-and-statements/category-1/video-sharing-platform-regulation.

Ofcom. *Ofcom's Strategy and Priorities for the Promotion of Media Literacy* (London: Ofcom, 2004). Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0021/72255/strat_prior_statement.pdf.

Ofcom. *Regulating video-sharing platforms A guide to the new requirements on VSPs and Ofcom's approach to regulation* (London: Ofcom, 2020). Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0021/205167/regulating-vsp-guide.pdf.

Oltermann, Philip. "Tough new German law puts tech firms and free speech in the spotlight", *The Guardian,* 5 January 2018. Available at: https://www.theguardian.com/world/2018/jan/05/tough-new-german-law-puts-tech-firms-and-free-speech-in-spotlight.

Orlando, Joanne. “Young people are exposed to more hate online during COVID. And it risks their health”, *The Conversation*, 9 November 2020. Available at: https://theconversation.com/young-people-are-exposed-to-more-hate-online-during-covid-and-it-risks-their-health-148107.

Ottoni, Raphael et al. “Analyzing right-wing YouTube channels: hate, violence and discrimination”, *Proceedings of the 10th ACM Conference on Web Science* (2018). Available at: https://arxiv.org/pdf/1804.04096.pdf.

Parekh, Bhikhu. “Is there a case for banning hate speech?”, pp. 37–56 in M. Herz and P. Molnar (eds.) *The content and context of hate speech: Rethinking regulation and responses* (Cambridge: Cambridge University Press, 2012).
Park, Ji Hoon et al. “Naturalizing Racial Differences Through Comedy: Asian, Black, and White Views on Racial Stereotypes in Rush Hour 2”, *Journal of Communication,* 56: 1, pp. 157-177 (2006). Available at: https://doi.org/10.1111/j.1460-2466.2006.00008.x.

Penney, Jonathon W. “Internet surveillance, regulation, and chilling effects online: a comparative case study”, *Internet Policy Review*, 6: 2, pp. 1-39 (2017). Available at: http://dx.doi.org/10.14763/2017.2.692.

Perez, Sarah. “Following riots alternative social apps and private messengers top the app stores”, *Tech Crunch*, 11 January 2021. Available at: https://techcrunch.com/2021/01/11/following-riots-alternative-social-apps-and-private-messengers-top-the-app-stores/.

Pressgrove, Geah et al. “What is Contagious? Exploring why content goes viral on Twitter: A case study of the ALS Ice Bucket Challenge”, *International Journal of Nonprofit and Voluntary Sector*, pp. 1-8 (2020). Available at: https://doi.org/10.1002/nvsm.1586.

r/TheDonald. Secondary website. Available at: https://thedonald.win. Last accessed on 4 December 2020.

r/TheRedPill. Secondary website. Available at: http://trp.red. Last accessed on 4 December 2020.

Rawlinson, Kevin. “Viscount who was jailed over Gina Miller threats drops his appeal”, *The Guardian*, 25 August 2017. Available at: https://www.theguardian.com/politics/2017/aug/25/viscount-jailed-gina-miller-threats-drops-appeal-sentence.

REACT, *National qualitative and quantitative report: United Kingdom* (Milan: React No Hate, 2018). Available at: http://www.reactnohate.eu/wp-content/uploads/2019/09/D2.3_REACT_UK-National-Qualitative-and-quantitative-Report-on-the-monitoring-results.pdf.

Read Across The Aisle. “Read Across the Aisle” (2020). Available at: http://www.readacrosstheaisle.com. Last accessed on 4 December 2020.

Reddit, “Promoting Hate Based on Identity or Vulnerability”. Available at: https://www.reddithelp.com/hc/en-us/articles/360045715951. Last accessed on 4 December 2020.

Reynders, Didier. *Countering Illegal Hate Speech Online - 5th Evaluation of the Code of Conduct* (Brussels: European Commission, 2020). Available at: https://ec.europa.eu/info/sites/info/files/codeofconduct_2020_factsheet_12.pdf.
Reynolds, Louise. “Defeating hate speech online”, *Institute for Strategic Dialogue.* Available at: https://www.isdglobal.org/defeating-hate-speech-online/. Last accessed on 4 December 2020.

Ribeiro, Manoel et al. “Auditing radicalization pathways on YouTube”, *Proceedings of the 2020 Conference on Fairness, Accountability and Transparency*, pp. 131-141 (2020). Available at: https://arxiv.org/pdf/1908.08313.pdf.

Roberts, Sarah T. “Content moderation”, in Larry Schintler and Clea McNeel (eds.), *Encyclopaedia of Big Data* (Berlin: Springer, 2017).
Roberts, Sarah T. *Behind the Screen: Content Moderation in the Shadows of Social Media* (New Haven; London: Yale University Press, 2019).

Rogers, Richard. “Deplatforming: Following Extreme Internet Celebrities to Telegram and Alternative Social Media”, *European Journal of Communication* 35: 3, pp. 213-229 (2020). Available at: https://doi.org/10.1177/0267323120922066.

Roose, Kevin. “The Making of a YouTube Radical”, *The New York Times*, 8 June 2019. Available at: https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html

Roozenbeek, Jon et al. “Susceptibility to misinformation about COVID-19 around the world”, *Royal Society Open Science*, 7: 10, pp. 1-15 (2020). Available at: https://royalsocietypublishing.org/doi/10.1098/rsos.201199.

Rossini, Patrícia. “Beyond Incivility: Understanding Patterns of Uncivil and Intolerant Discourse in Online Political Talk,” *Communication Research* (2020). Available at: https://doi.org/10.1177/0093650220921314.

Röttger, Paul et al. “HateCheck: Functional Tests for Hate Speech Detection”, *arXiv:2012.15606v1* (2020). Available at: https://arxiv.org/pdf/2012.15606.pdf.

Royal Courts of Justice. *The Queen on the application of Harry Miller v (1) The College of Policing and (2) The Chief Constable of Humberside CO/2507/2019* (London: Royal Courts of Justice, 2020). Available at: https://www.judiciary.uk/wp-content/uploads/2020/02/miller-v-college-of-police-judgment.pdf.

Runnymede Trust, *Islamophobia: Still a challenge for us all*, (London: The Runnymede Trust, 2017).

Salminen, Joni et al. "Online Hate Interpretation Varies by Country, But More by Individual: A Statistical Analysis Using Crowdsourced Ratings," in *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)* (Valencia: IEEE), pp. 88-94 (2018). Available at: https://ieeexplore.ieee.org/document/8554954.

Salminen, Joni et al. “Online hate ratings vary by extremes: a statistical analysis”, *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, pp. 213-217 (2019). Available at: https://doi.org/10.1145/3295750.3298954

Sap, Maarten et al. “The Risk of Racial Bias in Hate Speech Detection”, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence: Association for Computational Linguistics), pp. 1668–1678 (2019). Available at: https://www.aclweb.org/anthology/P19-1163.

Saunders, Alison. “Hate is hate. Online abusers must be dealt with harshly,” *The Guardian,* 21 August 2017. Available at: https://www.theguardian.com/commentisfree/2017/aug/20/hate-crimes-online-abusers-prosecutors-serious-crackdown-internet-face-to-face.

Shu, Catherine. “Cloudflare will stop service to 8chan, which CEO Matthew Prince describes as a ‘cesspool of hate’”, *Tech Crunch*, 5 August 2019. Available at: https://techcrunch.com/2019/08/04/cloudflare-will-stop-service-to-8chan-which-ceo-matthew-prince-describes-as-a-cesspool-of-hate/.

Skatova, Anya et al. “Unpacking Privacy: Willingness to Pay to Protect Personal Data,” *PsyArXiv* (2019). Available at: https://psyarxiv.com/ahwe4/.

Skopek, Jeffrey. “Anonymity, the Production of Goods and Institutional Design”, *Fordham Law Review*, 82: 4, pp. 1751-1809. Available at: https://ir.lawnet.fordham.edu/flr/vol82/iss4/4.

Sleeping Giants, https://www.slpnggiants.com/. Last accessed on 4 December 2020.

Snapchat, “Community Guidelines”. Available at: https://www.snap.com/en-GB/community-guidelines. Last accessed on 4 December 2020.

Statt, Nick. “YouTube decides that homophobic harassment does not violate its policies”, *The Verge*, 4 June 2019. Available at: https://www.theverge.com/2019/6/4/18653088/youtube-steven-crowder-carlos-maza-harassment-bullying-enforcement-verdict.

Stop Hate for Profit, https://www.stophateforprofit.org/. Last accessed on 4 December 2020.

Strossen, Nadine. *HATE: Why We Should Resist it with Free Speech, Not Censorship* (Oxford: Oxford University Press, 2018)

Sureka, Ashish et al. "Mining YouTube to Discover Extremist and Hidden Communities", *Lecture Notes In Computer Science: Asia Information Retrieval Symposium*, 6458, pp. 13-24 (2010). Available at: https://link.springer.com/chapter/10.1007/978-3-642-17187-1_2.

TeamYouTube (@TeamYouTube) Twitter post, "We came to this decision because a pattern of egregious actions has harmed the broader community and is against our YouTube Partner Program policies" (San Francisco: Twitter, 5 June 2019). Available at: https://twitter.com/teamyoutube/status/1136341801109843968?lang=en. Last accessed on 4 December 2020.

TeamYouTube (@TeamYouTube) Twitter post. "Our teams spent the last few days conducting an in-depth review of the videos flagged to us, and while we found language that was clearly hurtful, the videos as posted don't violate our policies" (San Francisco: Twitter, 5 June 2019). Available at: https://twitter.com/TeamYouTube/status/1136055351885815808. Last accessed on 4 December 2020.

TeamYouTube (@TeamYouTube) Twitter post. "Sorry for the confusion, we were responding to your tweets about the T-shirt" (San Francisco: Twitter, 5 June 2019). Available at: https://twitter.com/teamyoutube/status/1136363701882064896?s=21. Last accessed on 4 December 2020.

The Alan Turing Institute, *Response of the Public Policy Programme to the DCMS and the Home Office's Online Harms White Paper* (London: The Alan Turing Institute, 2019). Available at: https://www.turing.ac.uk/sites/default/files/2019-07/response_of_the_public_policy_programme_to_the_dcms_and_the_home_offices_online_harms_white_paper.pdf.

The Alan Turing Institute. “Hate Speech: Measures and Counter-Measures”. Available at: https://www.turing.ac.uk/research/research-projects/hate-speech-measures-and-counter-measures. Last accessed on 4 December 2020.

The Anti-Defamation League. “Best Practices” and Counterspeech Are Key to Combating Online Harassment”, 7 March 2016. Available at:https://www.adl.org/blog/best-practices-and-counterspeech-are-key-to-combating-online-harassment.

The Centre for Countering Digital Hate. Available at: https://www.counter-hate.com. Last accessed on 4 December 2020.

The NYU Dispatch. “Instagram vs TikTok: The Battle Between Social Media Platforms” (New York: The NYU Dispatch, 20 February 2020). Available at: https://wp.nyu.edu/dispatch/2020/02/20/instagram-vs-tiktok-the-battle-between-social-media-platforms/.

The Royal Society*. Machine learning: the power and promise of computers that learn by example* (London: The Royal Society, 2017). Available at: https://royalsociety.org/~/media/policy/projects/machinelearning/publications/machine-learning-report.pdf?la=en-GB&hash=B4BA640A1B3EFB81CE4F79D70B6BC234.
Tiffany, Kaitlyn. “My Little Pony Fans Are Ready to Admit They Have a Nazi Problem”, *The Atlantic*, 23 June 2020. Available at: https://www.theatlantic.com/technology/archive/2020/06/my-little-pony-nazi-4chan-black-lives-matter/613348/.

TikTok, “Community Guidelines”. Available at: https://www.tiktok.com/community-guidelines?lang=en. Last accessed on 4 December 2020.

TikTok. “Adding clarity to content removals”, 22 October 2020. Available at: https://newsroom.tiktok.com/en-us/adding-clarity-to-content-removals. Last accessed on 4 December 2020.

Titcomb, James. “Twitter blocks banned users from creating new accounts”, *The Telegraph*, 7 February 2017. Available at:

https://www.telegraph.co.uk/technology/2017/02/07/twitter-blocks-banned-users-creating-new-accounts/.

Tumblr. "Community Guidelines". Available at: https://www.tumblr.com/policy/en/community. Last accessed on 4 December 2020.

Tusikov, Natasha. "Defunding Hate: PayPal's Regulation of Hate Groups." *Surveillance & Society* 17, no. 1/2, pp. 46-53 (2019). Available at: https://ojs.library.queensu.ca/index.php/surveillance-and-society/article/view/12908/8475.

Twitch. "Hateful Conduct and Harassment". Available at: https://www.twitch.tv/p/en-gb/legal/community-guidelines/harassment/#hateful-conduct. Last accessed on 4 December 2020.

Twitter. "Hateful Conduct Policy". Available at: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy. Last accessed on 4 December 2020.
UK Government Press Release. "Government minded to appoint Ofcom as online harms regulator", 12 February 2020. Available at: https://www.gov.uk/government/news/government-minded-to-appoint-ofcom-as-online-harms-regulator
UK Legislation. *Public Order Act 1986* (London: UK,1986). Available at: https://www.legislation.gov.uk/ukpga/1986/64.

UK Legislation. *The Audiovisual Media Services Regulations 2020* (London: UK, 2020). Available at: https://www.legislation.gov.uk/uksi/2020/1062/made

UK Safer Internet Centre. "Government says new online harms legislation is expected to be ready next year" (London: UK Safer Internet Centre, 9 October 2020), Available at: https://www.saferinternet.org.uk/blog/government-says-new-online-harms-legislation-expected-be-ready-next-year.

Ullmann, Stefanie and Marcus Tomalin. "Quarantining online hate speech: technical and ethical perspectives", *Ethics and Information Technology*, 22:1, pp. 69-80 (2020). Available at: https://doi.org/10.1007/s10676-019-09516-z.

UN Office of the High Commissioner for Human Rights. "Governments And Internet Companies Fail To Meet Challenges Of Online Hate – UN Expert", 21 October 2019. Available at: https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=25174&LangID=E.

UN Office of the High Commissioner for Human Rights. *Annual report of the United Nations High Commissioner for Human Rights: addendum* (Geneva: OHCHR, 2013). Available at: https://www.ohchr.org/Documents/Issues/Opinion/SeminarRabat/Rabat_draft_outcome.pdf.

UN Office of the High Commissioner for Human Rights. *Rabat Threshold Test* (New York: OHCHR, 2020). Available at: https://www.ohchr.org/Documents/Issues/Opinion/Articles19-20/ThresholdTestTranslations/Rabat_threshold_test.pdf

United Nations. *United Nations Strategy and Plan of Action on Hate Speech* (New York: United Nations, 2019). Available at: https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf.

Vidgen, Bertie and Josh Cowls. "What is the harm in online hate?", *Forthcoming* (2021).

Vidgen, Bertie and Taha Yasseri. "Four ways social media platforms could stop the spread of hateful content in aftermath of terror attacks", *The Conversation*, 18 March 2020. Available at: https://theconversation.com/four-ways-social-media-platforms-could-stop-the-spread-of-hateful-content-in-aftermath-of-terror-attacks-113785.

Vidgen, Bertie et al. "Challenges and frontiers in abusive content detection" in *Proceedings of the Third Workshop on Abusive Language Online* (Florence: Association for Computational Linguistics), pp. 80-93 (2020). Available at: https://www.aclweb.org/anthology/W19-3509/.

Vidgen, Bertie et al. *An agenda for research into online hate* (London: The Alan Turing Institute, 2020). Available at: https://www.turing.ac.uk/research/publications/agenda-research-online-hate
Vidgen, Bertie et al. *How Much Online Abuse Is There? A Systematic Review of Evidence for the* UK (London: The Alan Turing Institute, 2019). Available at: https://www.turing.ac.uk/sites/default/files/2019-11/online_abuse_prevalence_full_24.11.2019_-_formatted_0.pdf.

Vidgen, Bertie et al.. "Recalibrating classifiers for interpretable abusive content detection classifiers", *Proceedings of the Fourth Workshop of the Natural Language Processing and Computational Social Sciences*, pp. 132-133 (2020). Available at: https://www.aclweb.org/anthology/2020.nlpcss-1.14.pdf.

Vidgen, Bertie. "Tweeting Islamophobia: Islamophobic hate speech amongst followers of UK political parties on Twitter – Doctoral thesis", (Oxford: University of Oxford, 2019). Available at: https://www.voxpol.eu/download/phd_thesis/Tweeting-Islamophobia-Islamophobic-hate-speech-amongst-followers-of-UK-political-parties-on-Twitter.pdf.

Vidgen, Bertie. Profile on The Alan Turing Institute website. Available at: https://www.turing.ac.uk/people/researchers/bertie-vidgen. Last accessed on 4 December 2020.

Vimeo. "Vimeo Acceptable Use Community Guidelines". Available at: https://vimeo.com/help/guidelines. Last accessed on 4 December 2020.

Vincent, James. "YouTube is deleting comments with two phrases that insult China's Communist Party", *The Verge*, 26 May 26 2020. Available at: https://www.theverge.com/2020/5/26/21270290/youtube-deleting-comments-censorship-chinese-communist-party-ccp.

Vincent, James. "YouTube says China-linked comment deletions weren't caused by outside parties", *The Verge*, 28 May 2020. Available at: https://www.theverge.com/2020/5/28/21272983/youtube-deleting-comments-chinese-communist-censorship-explanation.

Waldron, Jeremy. *The Harm in Hate Speech* (Oxford: Oxford University Press, 2012).

Waseem, Zeerak. “Are you racist or am I seeing things? Annotator influence on hate speech detection on Twitter”, *Proceedings of the 2016 EMNLP Workshop on Natural Language Processing and Computational Social Science*, pp. 138-142 (2016). Available at: https://www.aclweb.org/anthology/W16-5618

Weaver, Simon. “A rhetorical discourse analysis of online anti-Muslim and anti-Semitic jokes”, *Ethnic and Racial Studies*, 36: 3, pp. 483-499 (2011). Available at: https://www.tandfonline.com/doi/abs/10.1080/01419870.2013.734386.

Weber, Anne. *Manual on hate speech* (Strasbourg: Council of Europe Publishing, 2009)

Weimann, Gabriel and Natalie Masri. "Research Note: Spreading Hate On Tiktok", *Studies In Conflict & Terrorism*, pp. 1-14 (2020). Available at: https://www.tandfonline.com/doi/abs/10.1080/1057610X.2020.1780027.

White, Fiona, Lauren Harvey and Hisham Abu-Rayya. “Improving intergroup relations in the Internet age: a critical review”, *Review of General Psychology*, 19: 2, pp. 129-139 (2015) Available at: https://doi.org/10.1037/gpr0000036.

Williams, James. *Out of Our Light: Freedom and Resistance in the Attention Economy* (Cambridge: Cambridge University Press, 2018).

Williams, Matthew and Pete Burnap. “Cyberhate on Social Media in the aftermath of Woolwich: A Case Study in Computational Criminology and Big Data,” *The British Journal of Criminology*, 56: 2, pp. 211-238 (2016). Available at: https://doi.org/10.1093/bjc/azv059.

Williams, Matthew et al. "Hate In The Machine: Anti-Black And Anti-Muslim Social Media Posts As Predictors Of Offline Racially And Religiously Aggravated Crime", *The British Journal Of Criminology*, 60: 1, pp. 93-117 (2020). Available at: https://academic.oup.com/bjc/article/60/1/93/5537169.

Williams, Matthew. *Hatred Behind the Scenes* (London: Mishcon de Reya, 2020). Available at: https://hatelab.net/wp-content/uploads/2019/11/Hatred-Behind-the-Screens.pdf.

Winter, Aron. “Online Hate: From the Far-Right to the ‘Alt-Right’ and from the Margins to the Mainstream,” in *Online Othering. Palgrave Studies in Cybercrime and Cybersecurity*, Karen Lumsden and Emily Harmer eds. (London: Palgrave Macmillan, 2019). Available at: https://doi.org/10.1007/978-3-030-12633-9_2.

Wolf, Marie et al. “Why we should have seen this coming: Comments on Microsoft’s Tay ‘Experiment’ and Wider Implications”, *The ORBIT Journal*, 1: 2, pp. 1-12 (2012). Available at: https://doi.org/10.29297/orbit.v1i2.49

Xu, Jing et al. “Recipes for Safety in Open-domain Chatbots”, *Arxiv:2010.07079v2* (2020). Available at: https://arxiv.org/pdf/2010.07079.pdf.

YouTube. “Appeal Community Guidelines actions”. Available at: https://support.google.com/youtube/answer/185111. Last accessed on 4 December 2020.

YouTube. “Hate Speech Policy”. Available at: https://support.google.com/youtube/answer/2801939?hl=en-GB. Last accessed on 4 December 2020.

Zannettou, Savvas et al. “What Is Gab? A Bastion of Free Speech or an Alt-Right Echo Chamber?” in *Proceedings of the International World Wide Web Conference ACM* (Lyon: International World Wide Web Conferences), pp. 1007-1014 (2018). Available at: https://doi.org/10.1145/3184558.3191531.

turing.ac.uk
@turinginst