An agenda for research into online hate

Bertie Vidgen, Alex Harris,
Josh Cowls, Ella Guest,
Helen Margetts

The Alan Turing Institute

Public Policy Programme
Hate Speech: Measures
and Counter Measures

Executive summary

Online hate is a 'wicked problem' in the truest sense: 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. Good scientific research can help to address these wicked problems but, for too long, those on the frontlines in the fight against online hate (including civil society advocates, policy makers, regulators and politicians) have not fully benefited from academic research. This situation urgently needs to be rectified so that academic expertise is leveraged to better inform how online hate is tackled, its effects minimised and support provided to victims.

Through interviews and discussions with a range of stakeholders, as well as events and workshops, literature surveys and new empirical research, researchers at The Alan Turing Institute's Public Policy Programme have developed a six-point research agenda. This is intended as one step towards achieving the goal of policy-oriented and problem-driven academic research into online hate.

1. Online hate has serious and long-lasting impact on victims, their communities and societies at large. More research is needed into its effects.
2. Research into online hate often does not engage with the needs of society. It needs to be solution-driven and informed by the concerns and priorities of stakeholders.
3. Research into online hate needs to be flexible and responsive, balancing long-term studies with insights that have immediate impact
4. Online hate will always be a contentious area of research – definitions should be stated clearly, and all assumptions made explicit.
5. Data intensive technologies are not a silver bullet. If they are to be used, they must be used responsibly.
6. A positive vision of the Internet must be articulated and defended.

These agenda points lead us to three recommendations which we believe will foster the kind of high impact, solution-oriented research that is needed to address the growing problem of hate speech.

Recommendations

For researchers, collaborate! More pathways for collaboration across sectors need to be created, including opportunities for discussion and dialogue (e.g. workshops, conferences and roundtables) as well as joint working (e.g. shared research projects, co-authored publications and advisory roles). Collaborative relationships also need to be built within academia, bringing together researchers from social and computer sciences, as well as the humanities, through inter- and multi-disciplinary initiatives for researching online hate. Policy makers can play a unique role in enabling collaboration, identifying problems that hate speech causes or exacerbates and areas where statutory authority, such as regulation, is required. They can also help to join up research capacity across public agencies and minimise duplication of efforts. More broadly, the earlier that collaboration takes place between stakeholders, the more that gains can be realised.

For funders and policy makers, incentivise! Online hate research often involves crossing sectoral and disciplinary boundaries, involving researchers with different motivations and concerns. Targeted incentivisation is needed to assemble multi-disciplinary teams with the required social and computational expertise, and to conduct robust, impactful research at scale. To help achieve this, research councils should open up funding routes for academics who are working closely with community groups, policy makers and regulators. This is also a question of delivery and timing: some aspects of research are highly resource-intensive (such as developing privacy enhancing technologies (PETs) to enable research on sensitive data, and engineering new machine learning classification systems) – they need sustained multiyear funding. Once these resource-intensive opportunities have been set up, other research can be done efficiently on a smaller scale. Equally, small-scale ways of advancing research are through competitions, seed funding and incentivisation with impact – and data – opportunities.

For social media companies, open up! Social media companies are the gatekeepers of the data needed to do research, and will be the ones that ultimately design and implement interventions during the hate speech 'lifecycle'. Although some companies have made efforts to improve the transparency of their operations, with TikTok announcing in July 2020 that it aims to make its content moderation algorithms fully public[1], there are growing concerns that too much of the industry is opaque and inaccessible. It is crucial that companies continue to open up about their technologies, moderation processes, governance and data. Conversely, researchers at social media companies can benefit from research innovations, critical discussions and consensus-building around normative questions (such as defining hate) and policy understanding. One challenge is building such collaborations whilst keeping appropriate barriers in place between companies and regulators – this is essential if solutions are to be developed and put into practice.

1 The Verge. 2020. 'TIkTok is opening up its algorithms and challenging competitors to do the same.' The Verge, 29 July. Available at: https://www.theverge.com/2020/7/29/21346390/tiktok-algorithm-moderation-policy-transparency-china

Resources for researching online hate

The field of online hate research is expanding rapidly. To help navigate the field, we provide a non-exhaustive list of some useful resources:

- The Online Hate Research Hub: a new resource from The Alan Turing Institute, collating existing resources and providing links to other organisations and relevant outputs
- Hatespeechdata.com: a list of training datasets to create automated detection and classifications systems for online hate, curated by The Alan Turing Institute
- Turing Public Policy Briefing: How Much Online Abuse Is There? A systematic review of evidence for the UK
- SELMA, Hacking Hate: an EU-funded project that has created resources to tackle hate off all kinds
- The Centre for Countering Digital Hate, 'Don't Feed The Trolls' online campaign and guidance.
- International Data-driven Research for Advanced Modelling and Analysis (iDrama Lab): a group of researchers who conduct a wide range of research into online harms (alongside other phenomena)
- HateLab, Cardiff University: a global hub for research and insights into hate speech and crime
- Centre for Hate Studies, University of Leicester: a criminology centre conducting pioneering research on issues of hate and extremism

Introduction

The six-point research agenda presented here is intended as one step towards achieving the goal of policy-oriented and problem-driven academic research on online hate.[2] It has been developed through interviews and discussions with policymakers, regulators, civil society organisations, academics, think tanks, safety-tech companies and large social media platforms over the past two years. We have also hosted public-facing events[3] and closed workshops, reviewed a vast amount of academic literature[4], and conducted new empirical studies.[5] Inevitably, not every insight, contribution and piece of information could be incorporated, but the points raised here capture the most salient aspects of the contributions and feedback we received. We would like to thank everyone who fed into the creation of this agenda.

The health crisis caused by COVID-19 has had far-reaching consequences, not least by driving new and unanticipated forms of online hate and exposing the limitations of our existing responses.[6] The UN Secretary General has described a 'tsunami of hate and xenophobia'[7] being unleashed, and we echo his concerns about the dangerous social hazards being created. Equally, the Black Lives Matter movement and associated protests motivated by the tragic killing of George Floyd, has highlighted the divisions, inequalities and injustices which still exist around the globe, particularly in respect to the victims of racism. During these difficult times, support should be offered to all targets of online hate and our efforts to tackle such behaviour should be redoubled.

2 Shapiro, I., 2002. 'Problems, Methods, and Theories in the Study of Politics, or What's Wrong with Political Science and What to Do about It'. Political theory, 30(4), pp.596-619.
3 More information available at: https://www.turing.ac.uk/events/hate-and-harassment-can-technology-solve-online-abuse
4 Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S. and Margetts, H., 2019, August. 'Challenges and frontiers in abusive content detection'. Association for Computational Linguistics, 3rd workshop on abusive language online; Vidgen, B. and Derczynski, L., 2020. 'Directions in Abusive Language Training Data: Garbage In, Garbage Out'. arXiv preprint: 2004.01670.
5 Vidgen, B., Botelho, A., Broniatowski, D., Guest, E., Hall, M., Margetts, H., Tromble, R., Waseem, Z. and Hale, S., 2020. 'Detecting East Asian Prejudice on Social Media'. arXiv preprint:2005.03909.
6 Cowls, J., Vidgen, B. and Margetts, H. 2020. 'Why content moderators should be key workers', The Alan Turing Institute, 15 April. Available at: https://www.turing.ac.uk/blog/why-content-moderators-should-be-key-workers
7 CBS News. 2020. 'UN Chief says pandemic spurring tsunami of hate'. CBS News, 8 May. Available at: https://www.cbsnews.com/news/un-chief-says-pandemic-spurring-tsunami-of-hate

Agenda

1. Online hate has serious and long-lasting impact on victims, their communities and societies at large. More research is needed into its effects.

Online hate can inflict harm in a variety of ways, directly affecting individuals and groups who are targeted, as well as the communities they are from and others who may be exposed to the hate (even if it is not directed at them). These harms can easily seep from the online to the offline, causing mental health problems and anxiety and even making people fear leaving their homes.[8] Online hate has also been connected with offline violence and attacks.[9] The brunt of online hate and harassment often falls on already-marginalised and under-represented communities, including those who sit at the intersection of multiple identities which make them especially vulnerable. The huge disparities in which groups are most impacted by online hate is directly linked to broader social, historical and cultural contexts. Centuries of racial prejudice against Black people, for example - much of it legally institutionalised in the form of slavery and apartheid – have resulted in persistent racist structures that continue to impose disproportionate burdens on Black people.[10] Similar inequalities and structural challenges are faced by other non-majority groups across the UK.

Many advocacy groups are doing important work to understand the impact of online hate on targeted groups, including Tell Mama (Islamophobia), Stonewall (prejudice against LGBT people), the CST (antisemitism), Glitch (women) and the Centre for Countering Digital Hate (multiple). But more research is needed to better understand the impact of online hate on victims, their communities, social media platforms and wider society – and, more broadly, academics need to put those who are affected by online hate at the forefront of all research. Further, meaningful ties should be forged between academics and advocates, helping to ensure that victims' voices are heard, their priorities acted upon, and their perspectives incorporated into research design. Researchers should also think about how insights can be collected from communities in a responsible and respectful way that is not purely extractive; participatory research designs are an excellent avenue for this.

8 Gelber, K. and McNamara, L. 2016. 'Evidencing the harms of hate speech'. Social Identities, 22(3), pp.324-341.
9 Williams, M., Burnap, P., Javed, A., Liu, H. and Ozalp, S. (2020). 'Hate in the Machine: anti-black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime', 60(1), pp. 93-117.
10 Modood, T., Werbner, P. and Werbner, P.J. eds. 1997. The Politics of Multiculturalism in the New Europe: Racism, Identity and Community. Palgrave Macmillan; Virdee, S. and McGeever, B. 2018. 'Racism, crisis, Brexit'. Ethnic and Racial Studies, 41(10), pp.1802-1819.

2. Research into online hate often does not engage with the needs of society. It needs to be solution-driven and informed by the concerns and priorities of stakeholders.

Without risking their impartiality, academic research on online hate could benefit by engaging more closely with the concerns and priorities of the stakeholders who use research. One challenge is that, by its nature, academic research tends to be critical and reflective, generating deep understanding about problems. Far less attention is paid to creating products, tools and frameworks which would solve problems. Note that 'solutions' includes more than just computational tools: activist toolkits[11], workshops[12], practical advice[13] and digital citizenship materials[14] are all research artefacts which help in tackling online hate. More collaboration is crucial to ensuring that useful solutions are actually developed. In this regard, we note some existing collaborations that have broadly been successful in producing impactful research. For instance, a research call organised by the Commission for Countering Extremism in 2018/2019 shows how academic-policymaker collaborations can be successfully organised, leading to 19 papers published.[15] Equally, the House of Commons' All-Party Parliamentary Groups[16] and select committee hearings[17] and reports[18] have been effective at bringing academics into direct dialogue with politicians.

Part of the challenge is that academics do not always know what issues need to be addressed. The Department for Digital, Culture, Media and Support (DCMS) has provided a list of research interests, which are a helpful guide for online harms research more generally.[19]

11 Glitch, Toolkit, < https://fixtheglitch.org/glitchukresources/toolkit>
12 Glitch, Workshops, <https://fixtheglitch.org/workshops/>
13 Centre for Countering Digital Hate, <https://www.counterhate.co.uk>
14 Reynolds, L. and Scott, R. 2016. 'Digital Citizens: Countering Extremism Online", Demos.
15 Commission for Countering Extremism, 2019, Call for evidence,. Available at: https://www.gov.uk/government/consultations/extremism-in-england-and-wales-call-for-evidence.
16 For instance: https://appgbritishmuslims.org and http://www.appghatecrime.org.
17 For instance: https://www.parliament.uk/business/committees/committees-a-z/commons-select/digital-culture-media-and-sport-committee/sub-committee-on-disinformation/
18 For instance: https://publications.parliament.uk/pa/cm201617/cmselect/cmhaff/609/609.pdf
19 The Department for Digital, Culture, Media and sport, 'DCMS Areas of Research Interest'. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/707598/DCMS_ARI_2018.pdf.

Beyond this, our reviews indicate some key areas in online hate research for which more evidence is needed:

- **The prevalence of online hate.** The Alan Turing Institute has conducted an initial review, but more evidence is still needed.[20]
- **Who is targeted by online hate**. Understanding the most frequent targets of hate is not only an important empirical issue, it is crucial for understanding the injustices of online hate, which are likely to disproportionately affect certain groups.[21]
- **The dynamics and patterns of online hate**. This should build understanding of when hate peaks and troughs, the role of networks and the role of platforms.
- **The global ecosystem of hate**. Understanding how hate intersects with other online harms, such as extremism and misinformation, and how it spreads across platforms, is a hard-to-study area that is essential for comprehensively understanding how hate manifests online.
- **The causes of online hate.** Causal analysis is notoriously difficult in social science, especially in settings where experiments are difficult to implement. This is particularly true with research into online hate where there are likely to be many biases in self-reporting.
- **Who sends hate and the degree to which hate is coordinated**. Whether hate is sent by actors operating in concert or by unconnected individuals behaving spontaneously remains an open question.[22] It likely depends on the setting and the actors involved. The Alan Turing Institute has proposed an initial framework to help understand how individuals engage in harmful behaviour, but more evidence is still needed.[23]
- **The impact of online hate**. As discussed above, hate can impact victims, communities and wider society in myriad ways. This needs to be investigated further.

20 The Alan Turing Institute, 2019, Online Hate Monitor. Available at: https://www.turing.ac.uk/research/research-programmes/public-policy/online-hate-monitor.

21 Ghanea, N. 2013. 'Intersectionality and the Spectrum of Racist Hate Speech: Proposals to the UN Committee on the Elimination of Racial Discrimination'. Human Rights Quarterly, 35, pp.935.

22 Mariconti, E., Suarez-Tangil, G., Blackburn, J., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Serrano, J.L. and Stringhini, G. 2018. ''You Know What to Do': Proactive Detection of YouTube Videos Targeted by Coordinated Hate Attacks'. arXiv preprint:1805.08168.

23 Maple, C., Vidgen, B. and Margetts, H. 2020. 'Why online harms research needs new collaboration, direction and a shared sense of purpose", The Alan Turing Institute, 22 June. Available at: https://www.turing.ac.uk/blog/why-online-harms-research-urgently-needs-new-collaboration-direction-and-shared-sense-purpose.

3. Research into online hate needs to be flexible and responsive, balancing long-term studies with insights that have immediate impact.

The landscape of online hate is in constant flux; new platforms, influencers, types of language and tropes emerge quickly, even if they only last for a short while. Changes are unpredictable, often emerging in response to unanticipated political and social events such as terror attacks, protests and international events – and impacted by behind-the-scenes coordination between online activists and the activities of bots. This is problematic for academic research because (1) findings may fast become irrelevant and (2) even methods can go out of date.

Research designs need be matched with their aims. In-depth long-term studies using highly curated datasets are important for building deep theoretical knowledge and reflecting critically on online hate. However, these need to be complemented by responsive studies which are more exploratory and investigate emerging trends in a timely and digestible way. Initiatives like The Alan Turing Institute's public policy briefing series and reports produced by think tanks can help to achieve this.[24]

Relatedly, academic research in online hate needs to be more open: many publications do not share the datasets, code and guidelines they use. This not only undermines the integrity of research because it means it is not reproducible, it also limits the use of findings by non-academics. In effect, a huge amount of resources for online hate research are lost because they are not open source.

24 For example: see reports by Demos (https://demosuk.wpengine.com/wp-content/uploads/2018/08/PatternsOfHateCrimeReport-.pdf) and the Institute for Strategic Dialogue (https://institute.global/policy/designating-hate-new-policy-responses-stop-hate-crime).

4. Online hate will always be a contentious area of research – definitions should be stated clearly, and all assumptions made explicit.

Defining online hate remains tricky.[25] 'Hate' is a fundamentally contested and contentious concept, and full consensus is unlikely to ever be reached on where exactly the line should fall between hate and non-hate (or 'legitimate critique', as it has been called).[26] Given this, applied research must (a) provide a definition of hate and (b) be clear about what the definition entails. Intrinsically, this is a question of research validity: a definition that is suitable for studying overt forms of hate might not be suitable for studying more subtle varieties. Equally, assumptions made for one type of research might not be suitable for another. Such issues are also inherently normative as definitions are likely, in turn, to impact and reflect broader social and ethical concerns. These issues should be laid out clearly. To help future researchers, we point to several working definitions for online hate which are suitable for empirical analyses:

- The Commission for Countering Extremism offers a comprehensive account of 'hateful extremism', which includes 'behaviours that can incite and amplify hate, or engage in persistent hatred, or equivocate about and make the moral case for violence.[27]
- Ofcom, the UK's communication regulator, has Section 3 guidelines, which covers materials that could incite crime or disorder (including hatred and abuse) and derogatory, dehumanising, and threatening statements. This has been developed for analogue services (e.g. Radio and TV).[28]
- The Law Commission, the UK's statutory independent body for reviewing the law, has conducted (and started to publish) numerous reports into online abuse and hate, providing a thorough overview and critique of the existing law.[29] These reports provide clarity on the limits and shortfalls of current legal frameworks.

25 Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S. and Margetts, H., 2019, August. 'Challenges and frontiers in abusive content detection'. Association for Computational Linguistics.
26 Imhoff, R. and Recker, J. 2012. 'Differentiating Islamophobia: Introducing a new scale to measure Islamoprejudice and secular Islam critique'. Political Psychology, 33(6), pp.811-82
27 Commission for Countering Extremism. 2019. 'Challenging Hateful Extremism'. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/874101/200320_Challenging_Hateful_Extremism.pdf
28 Ofcom. 2019. 'The Ofcom Broadcasting Code. Section three; Crime, Disorder, Hatred and Abuse'. Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0021/132078/Broadcast-Code-Section-3.pdf
29 Law Commission, 2018, Abusive and Offensive Online Communications, https://www.lawcom.gov.uk/abusive-and-offensive-online-communications

5. Data intensive technologies are not a silver bullet. If they are to be used, they must be used responsibly.

Data-intensive technologies such as machine learning, natural language processing, and artificial intelligence are now widely used to identify, study and moderate online hate.[30] These technologies offer unmatched scalability and are the only feasible way of monitoring the hundreds of millions of posts made on the major platforms each day.[31] However, they also pose several limitations, which everyone who uses and works with such technology should be aware of.[32] Part of the challenge is that these 'big' computational methods are being applied to what has historically been viewed as a 'small data' problem. Hate is inherently complex and manifests online in unusual ways, often through shortcuts, obfuscation and dog whistles; these nuanced forms of hate remain challenging even for state-of-the-art classification systems. Technology can also introduce various forms of bias and unfairness, from having higher error rates for content produced by certain groups to performing worse on hate directed against some identities (and therefore providing them less protection).[33] Such technologies may also be used in inappropriate and crude ways.

All data-intensive technologies need to be used ethically and responsibly in online hate research. Ethical principles developed by The Alan Turing Institute for the public sector provide a useful starting point.[34]

30 Cambridge Consultants, 2019, 'Use of AI in Online Content Moderation'. Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf

31 Department of Digital, Culture, Media & Sport and the Home Office. 2019. 'Online Harms White Paper'. Available at: https://www.gov.uk/government/consultations/online-harms-white-paper

32 Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S. and Margetts, H., 2019, August. 'Challenges and frontiers in abusive content detection'. Association for Computational Linguistics.

33 Davidson, T., Bhattacharya, D. and Weber, I. 2019. 'Racial bias in hate speech and abusive language detection datasets'. arXiv preprint:1905.12516.

34 Leslie, D. 'Understanding Artificial Intelligence ethics and safety: a guide for responsible design and implementation of AI systems in the public sector', London: The Alan Turing Institute. Available at: https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf

6. A positive vision of the Internet must be articulated and defended.

Hate, as well as harassment and extremism, are ugly and undesirable parts of the Internet, and form part of a wider set of online harms which urgently need to be challenged and removed.[35] Whilst devoting time and resources to tackle these issues, a positive vision of the Internet needs to be advocated for, highlighting its constructive uses and pro-social value.[36] One example of this is the 'Good Web' project from Demos, which aims to measure and build public support for a version of the Internet that empowers and advances democracy and realises public good.[37] This is particularly important when, as COVID-19 has made abundantly clear, the Internet is not something that people can just 'opt out of'. It plays a fundamental role in how everyone works, communicates, socialises, stays entertained, finds information and engages politically. Ultimately, the Internet is a battleground where competing values, ideas and viewpoints play out – and to win that battle we need to make sure that as well as removing harm we also foreground and encourage the good – leading to online spaces that are accessible, just and fair.

35 Davidson, T., Bhattacharya, D. and Weber, I. 2019. 'Racial bias in hate speech and abusive language detection datasets'. arXiv preprint:1905.12516; Garg, S., Perot, V., Limtiaco, N., Taly, A., Chi, E.H. and Beutel, A. 2019. 'Counterfactual fairness in text classification through robustness'. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 219-226.

36 Margetts, H. 2016. 'Don't Shoot the Messenger! What part did social media play in 2016 US Election?', The Oxford Internet Institute, 15 November. Available at: https://www.oii.ox.ac.uk/blog/dont-shoot-the-messenger-what-part-did-social-media-play-in-2016-us-e%C2%ADlection/

37 Demos, 'The Good Web Project'. Available at: https://demos.co.uk/project/the-good-web-project/

An agenda for research into online hate

The Alan Turing Institute

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