collective wellbeing Carnegie UK

Ad hoc advice from Carnegie UK to United Nations Special Rapporteur on Minority Issues

concerning guidelines on combatting hate speech targeting minorities in social media.

Authors:
Professor Lorna Woods, School of Law, University of Essex, UK
William Perrin, Trustee, Carnegie UK
Maeve Walsh, Associate, Carnegie UK

ISBN: 978-1-912908-76-9

This advice responds to a request for support from Fernand de Varennes, the Special Rapporteur on Minority Issues, in drafting guidelines on combatting hate speech targeting minorities in social media made in his "Thematic Report: hate speech, social media and minorities" to the United Nations Human Rights Council of March 2021.[1]

Note: This work is based upon a draft code of practice on hate crimes on social media prepared in consultation with a wide range of civil society groups representing victims of hate crime in the United Kingdom by Carnegie UK.

1 "The Special Rapporteur invites States, the United Nations and its entities, and in particular OHCHR, to initiate a process to develop a global voluntary code of conduct for social media platforms to combat hate speech. He also requests that they support his efforts to draft guidelines on combatting hate speech targeting minorities in social media, as a matter of urgency, in 2021–2022." Special Rapporteur on Minority Issues, 'Thematic Report: hate speech, social media and minorities' A/HRC/46/57 HRC 46th Session 3 March 2021 https://www.ohchr.org/Documents/Issues/Minorities/SR/A_HRC_46_57.docx

Introduction - combatting hate speech towards minorities on social media

1. As the Special Rapporteur for Minority Issues de Varennes observes,[2] there are adverse human rights impacts for minorities arising from the operation of social media which facilitates hate speech ranging from low abuse and slurs to incitement to genocide. Addressing this problem is urgent and is likely to become more so as companies seek to create virtual environments. Despite increased attention to the issue, social media service providers have shown themselves unable or unwilling to take sufficient action.[3] The recent evidence from Facebook whistleblower, Frances Haugen, demonstrates this point clearly; and the ongoing nature of the problem is suggested by the fact that she is not the first such whistleblower. The United Nations Guiding Principles on Business and Human Rights[4] (UNGPs) are not being well implemented by social media companies, if implemented at all. While the B-Tech work[5] of the UN Office of the High Commissioner for Human Rights (OHCHR) is helpful, as the Secretary General noted, guidance from the UNHCHR has not yet tackled the specific issue of social media and hate speech.[6]

2. In response to the Special Rapporteur's urgent request for support, in this document Carnegie UK identifies and discusses considerations for Guidelines for social media companies on combatting hate speech and in Annex A offer some tentative draft Guidelines for social media companies to combat hate speech. We seek to demonstrate the principled pragmatism of former Special Representative Ruggie's approach[7]. We discuss how to tackle an acute problem through specific guidance, informed by victims. This report and draft Guidelines are set within the framework of generalised guidance on business conduct and human rights, conforming with the UNGPs and OECD guidance. We identify relevant links and commonalities with

2 Report of the Special Rapporteur on minority issues, 'A contextualization: a pandemic of hate' A/HRC/46/57 HRC 46th Session 3 March 2021, paragraphs 35-44 https://www.ohchr.org/Documents/Issues/Minorities/SR/A_HRC_46_57.docx

3 "Only five of the top 50 online content-sharing services issue transparency reports specifically about (TVEC) terrorist and violent extremist content#"' OECD Digital Economy Paper No. 296, 'Current Approaches to Terrorist and Violent Extremist Content among the Global Top 50 Online Content-Sharing Services', August 2020, p4 available: https://www.oecd-ilibrary.org/science-and-technology/current-approaches-to-terrorist-and-violent-extremist-content-among-the-global-top-50-online-content-sharing-services_68058b95-en and 'Human Rights Impact Assessment Facebook in Myanmar' Business and Social Responsibility October 2018, available: https://about.fb.com/wp-content/uploads/2018/11/bsr-facebook-myanmar-hria_final.pdf

4 "The UNGPs are the global standard for preventing and addressing the risk of adverse impacts on human rights linked to business activity, and they provide the internationally-accepted framework for enhancing standards and practices with regard to business and human rights." The 2011 text of the Guiding Principles can be found in HR/PUB/11/04, available: https://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf.

5 The B-Tech Project provides authoritative guidance and resources for implementing the United Nations Guiding Principles on Business and Human rights (UNGPs) in the technology space, available: https://www.ohchr.org/EN/Issues/Business/Pages/B-TechProject.aspx.

6 As noted in the "UN Secretary General's Roadmap for digital cooperation: implementation of the recommendations of the High Level Panel", Report of the Secretary-General (A/74/821), available: https://undocs.org/A/74/821, para 86.

7 Ruggie, John G., and Tamaryn Nelson, "Human Rights and the OECD Guidelines for Multinational Enterprises: Normative Innovations and Implementation Challenges." May 2015, available: https://www.hks.harvard.edu/sites/default/files/centers/mrcbg/programs/cri/files/workingpaper.66.oecd.pdf

existing general guidance such as the UNGPs, the UNGPs Interpretative Guide[8], the UN B-Tech project and in some cases the OECD Guidance on Responsible Business Conduct[9]. We also draw upon or refer to other work on Human Rights Impact Assessments by the World Bank Group and new work on HRIAs for technology businesses from the Danish Institute for Human Rights[10] and Essex Human Rights Centre. This advice is based upon our work on a code of practice on hate speech and social media that Carnegie compiled in consultation with a range of groups representing victims of hate speech on social media in the United Kingdom.[11] We acknowledge, in particular, the role of the Antisemitism Policy Trust in that original code. Carnegie's wide-ranging work on social media regulation also informs this advice document.[12]

Definitions and Scope of work

3. In this paper, we follow the approach of the UN Strategy and Plan of Action on Hate Speech, and its Detailed Guidance in defining 'hate speech' as:

> *"any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factors."*[13]

4 This report concerns online platforms, with specific reference to social media, but not to all internet intermediaries and information society services (ISS). Broader intermediaries and ISS, while important to the functioning of the Internet, do not have the same level of interconnectedness with content choice and behaviour as social media. The key aspect of social media platforms is that they facilitate user interaction and engagement between users. This could include online gaming platforms and virtual realities, as well as more obvious candidates such as Facebook, YouTube, Twitter, TikTok, SnapChat, and Instagram. The potential for multiway engagement distinguishes these platforms from private communications – essentially one-to-one.[14] The issue of whether a platform is encrypted or not is not decisive: Whatsapp,

8 OHCHR, "The Corporate Responsibility to Respect Human Rights: An Interpretive Guide", HR/PUB/12/02 available: https://www.ohchr.org/Documents/Publications/HR.PUB.12.2_En.pdf [accessed 4 Nov 2021].

9 OECD Due Diligence Guidance for Responsible Business Conduct, 2018, available: https://www.oecd.org/investment/due-diligence-guidance-for-responsible-business-conduct.htm [Accessed 22 July 2021].

10 Emil Lindblad Kernell and Cathrine Bloch Veiberg, *Guidance on Human Rights Impact Assessment of Digital Activities*, 23 November 2020 (Danish Institute for Human Rights and Essex Human Rights Centre), available: https://www.humanrights.dk/publications/human-rights-impact-assessment-digital-activities [accessed 22 July 2021].

11 Original UK version published 15 June 2021: 'DRAFT Code of Practice in respect of Hate Crime and wider legal harms.' This draft code of practice was developed with the input of the following organisations: Antisemitism Policy Trust, The Bishop of Oxford's Office, Glitch, Centenary Action Group, Faith Matters, Galop, Hope Not Hate, Institute for Strategic Dialogue, The Alan Turing Institute. It was later discussed at a workshop including those organisations, along with other civil society representatives and attendees from regulators and the major tech platforms in February 2021, available: https://www.carnegieuktrust.org.uk/blog/draft-code-of-practice-in-respect-of-hate-crime-and-wider-legal-harms/

12 All our work can be found here: https://www.carnegieuktrust.org.uk/project/harm-reduction-in-social-media

13 United Nations Strategy and Plan of Action on Hate Speech (May 2019) and its Detailed Guidance (2020), both available: https://www.un.org/en/genocideprevention/hate-speech-strategy.shtml.

14 In our view the issue of encryption is not determinative for the question of whether private communications are an issue.

which is reasonably well encrypted, allows for large groups facilitates multiway communications. The approach proposed could be adapted to apply to search engines[15].

Carnegie UK systemic approach to reducing hate speech

5. Carnegie UK's approach focusses on the systems that make up the social media platform and not directly on the content posted by users. The Carnegie approach recognises that the platforms, as well as being in a gatekeeper role, are not neutral as to how people discover and create content. Choices made by the platforms about how they design their services affect the content seen (e.g. default to autoplay, curated playlists, data voids and algorithmic promotion) and even produced (e.g. through financial incentives for content creators, or the feedback loop created through metrification; emojis create a new shorthand for communication[16]).

6. These design choices exacerbate or even exploit disinhibited behaviours[17] that arise in the online environment, whereby we make decisions that would be less likely in offline environments. This may have particular salience in the context of hate speech. This is not to say that users are automatons, or are automatically criminalised by the online environment, but rather than they may be nudged towards certain behaviours. "Frictionless communication" may lend itself to "fast thinking", rather than "slow"[18], potentially favouring emotive content with an emphasis on negative emotions. Of course, it may also be that these systems are open to manipulation, for example where an inflated impression of grass root support for an issue (eg anti-minority rhetoric[19]) is given, and the platform operators have not given sufficient thought to protecting the systems from manipulation and abuse[20]. This sort of abuse may be part of attitudes and behaviours generally and not purely an online phenomenon. Nonetheless, social media platforms may add to the problems and enable or facilitate abuse. Design choices and the provision of safeguards are important.

7. The hazards caused by such failures are not evenly experienced. The risks of technology and its affordances cannot be adequately assessed without taking into account the broader context. A number of reports by other Special Rapporteurs have noted, for example, the problems faced by women and what the Special Rapporteur for Freedom of Expression termed 'gendered censorship'.[21] The Special Rapporteur

15 In the United Kingdom draft Online Safety Bill (CP 405) the UK government applies an adapted subset of its proposals to search engines, available: https://www.gov.uk/government/publications/draft-online-safety-bill [accessed 22 July 2021].

16 Anne Wagner, Sarah Marusek and Wei Yu 'Sarcasm, the smiling poop, and E-discourse aggressiveness; getting far too emotional with emojis' (2020) 30 *Social Semiotics* 305 DOI: https://doi.org/10.1080/10350330.2020.1731151; there are additional issues around differential understanding of emojis potentially exacerbated by different 'fonts' used by different platforms.

17 Suler, John, 'The Online Disinhibition Effect. Cyberpsychology & behavior: the impact of the Internet, multimedia and virtual reality on behavior and society', (2004). 7. 321-6.

18 Kahneman, D., *Thinking, Fast and Slow* (London: Penguin Books, 2012).

19 See e.g. Institute for Strategic Dialogue, T*he networks and narratives of anti-refugee disinformation in Europe*, 1 July 2021, available: https://www.isdglobal.org/isd-publications/the-networks-and-narratives-of-anti-refugee-disinformation-in-europe/ [accessed 21 July 2021

20 See e.g. Samantha Bradshaw, Hannah Bailey and Philip N. Howard, I*ndustrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation*, available: https://demtech.oii.ox.ac.uk/research/posts/industrialized-disinformation/ [accessed 21 July 2021].

21 Report of the Special Rapporteur on Freedom of Expression (A76/258) 30 July 2021 [accessed 22 September 2021], para 12.

on Violence against Women noted:

> Women are both disproportionately targeted by online violence and suffer disproportionately serious consequences as a result. Their access to technology is also affected by intersectional forms of discrimination based on a number of other factors, such as race, ethnicity, caste, sexual orientation, gender identity and expression, abilities, age, class, income, culture, religion, and urban or rural setting.[22]

8. Focussing on platform systems and processes allows a greater range of possible interventions that are human rights compliant. Systems-based interventions may allow potentially conflicting human rights of the many platform users to be more optimally balanced than would be the case in a regime in which the only response is to take content down[23], as the Special Rapporteur on freedom of opinion and expression has noted.[24] More recently, Irene Khan suggested that it may be appropriate to use measures such as downranking, demonetizing, friction, warnings, geoblocking and counter-messaging.[25]

9. In Carnegie's systems-based approach 'system' has a double meaning. First, it refers to the software and business systems, and the fact that they are the focus of attention under this approach. While questions of content inevitably arise, they are dealt with indirectly. The mechanisms the Special Rapporteur highlights constitute a systems-based approach in which the design and functionalities of the platform are central. Such an approach does not, however, displace content rules. There are systems concerns here too. A service provider may have a policy prohibiting hate speech, but it might choose to run the platform in such a way that the policy is not enforced effectively: a weak system undermines the policy.

10. Secondly, the approach requires each business to introduce a system for risk assessment, risk mitigation and reparation. This challenges companies which seek to operate on the basis of 'naive innovation' or wilful blindness. The recent *Wall Street Journal* reporting reveals documents demonstrating that senior management seemingly chose to ignore issues flagged by employees; this reporting supports earlier claims by civil society actors.[26]

11. Social media companies create synthetic environments for their users based upon the service provider's software and business process decisions. Within these

22 Report of the Special Rapporteur on Violence against Women, (A/HRC/38/47), 18 June 2018, para 28, available: https://undocs.org/A/HRC/38/47 [accessed 21 September 2021].

23 L. Woods, The Carnegie Statutory Duty of Care and Fundamental freedoms, December 2019, available: https://d1ssu070pg2v9i.cloudfront.net/pex/pex_carnegie2021/2019/12/05125454/The-Carnegie-Statutory-Duty-of-Care-and-Fundamental-Freedoms.pdf [accessed 21 September 2021].

24 Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, (A/74/486), 19 October 2019, para 51, available: https://www.undocs.org/A/74/486 [Accessed 22 July 2021].

25 Irene Khan, Public Comment by UN Special Rapporteur on Freedom of Opinion and Expression Irene Khan on Facebook Oversight Board Case no. 2021-009, 9 September 2021, available: https://www.ohchr.org/Documents/Issues/Opinion/Legislation/Case_2021_009-FB-UA.pdf [accessed 21 September 2021]

26 See e.g. Center for Countering Digital Hate, Malgorithm: how Instagram's Algorithm Publishes Misinformation and Hate to Millions during a Pandemic, available: https://www.counterhate.com/malgorithm [accessed 21 September 2021].

environments, companies make choices about what is allowed to happen for good or ill. Given this role in creating the forum and its rules and thus in creating, facilitating or exacerbating problems, companies are well-placed to manage them. Moreover, making the service provider responsible for implementing better systems is economically efficient, consistent with the "polluter pays" principle[27] returning external costs to society into the service provider's production decision.

12. A system-focussed approach lends itself to due diligence in risk assessment as applied regulatory practice across a range of industries. Risk-based due diligence underpins the UNGPs,[28] the OECD Due Diligence Guidance for Responsible Business Conduct[29] (which is set against the UNGP framework) and the UNHCHR B-Tech guidance. The definition from the UN Strategy and Plan of Action on Hate Speech and the methodology of the Carnegie UK systemic approach are a foundation for tackling hate speech on social media. We discuss below thematic components of an overall approach beginning with the most significant: effective corporate risk assessment.

Responsibility, Risk Assessment, Mitigation and Remediation

Responsibility

13. Only active leadership in social media companies will combat hate speech arising on their respective services. It will require the service provider to spend time and money taking active steps to combat hate speech. This requires leadership from the top in the form of a clear policy statement.[30] The provider should be clear about its own values, including a clear recognition of the importance of all human rights, which are universal, indivisible, interdependent and interrelated.[31] The company governance structure should clearly allocate and delineate roles and responsibilities, with a clear route for reporting on concerns to be considered by senior management.

Risk Assessments and Risk Mitigation

14. There is a wealth of high-level guidance on risk assessment that social media

27 OECD, "Recommendation of the Council on the Implementation of the Polluter Pays Principle", 1974, available: https://legalinstruments.oecd.org/en/instruments/11.

28 Due diligence has been defined as "such a measure of prudence, activity, or assiduity, as is properly to be expected fro, and ordinarily exercised by, a reasonable and prudent [person] under the particular circumstances; not measured by any absolute standard, but depending on the relative facts of the special case". In the context of the Guiding Principles, human rights due diligence comprises an ongoing management process that a reasonable and prudent enterprise needs to undertake, in the light of its circumstances (including sector, operating context, size and similar factors) to meet its responsibility to respect human rights. OHCHR, ' Interpretative Guide' (n 8), p 6.

29 "The OECD Due Diligence Guidance for Responsible Business Conduct provides practical support to enterprises on the implementation of the OECD Guidelines for Multinational Enterprises by providing plain-language explanations of its due diligence recommendations and associated provisions." OECD Responsible Business Conduct (n 9).

30 See UNGP15 (n 4): 'In order to meet their responsibility to respect human rights, business enterprises should have in place policies and processes appropriate to their size and circumstances, including … a policy commitment to meet their responsibility to respect human rights'.

31 World Conference on Human Rights, Vienna Declaration and Programme of Action, 25 June 1993, (A/CONF 157/23), para 5.

companies do not appear to be following[32]. UNGPs and other such guidance provide high-level support to all companies for managing and averting the risk of human rights impacts. UNGP specifies that companies should have:

> *'A human rights due diligence process to identify, prevent, mitigate and account for how they address their impacts on human rights.'*[33]

15. The OECD Guidance on due diligence for responsible business conduct[34] provides a good framework, as does ISO 310000[35]. The basic principle is simple. As the OECD guidance[36] notes:

> *'Due diligence is risk-based. The measures that an enterprise takes to conduct due diligence should be commensurate to the severity and likelihood of the adverse impact. When the likelihood and severity of an adverse impact is high, then due diligence should be more extensive.'*

16. This breaks down into a number of aspects: define risk, understand the consequences; evaluate the likelihood; identify how the organisation could eliminate, mitigate, control or react to the risk; test and evaluate control measures; identify where improvement is needed. When identifying risk and control measures the differential impact on sub-sets of the user group should be taken properly into account.

17. Due diligence should be understood in the context of the business sector in issue, and companies should be aware of any sector specific standards too.[37] For the tech sector in general, the UNHRC B-Tech project[38] makes clear that this includes:

> *'a company identifying whether and how the design, development, promotion, deployment and use of its products and services could lead to adverse human rights impacts.'*

18. Focussing this recognition on social media brings into scope product design, business processes, community standards and moderation standards.[39] Cyber security is also relevant as social media accounts may be hacked, hijacked or faked to disastrous effect. The leaking of data (e.g. geolocation data inadvertently included by a user in a post) is also problematic.

32 We note that Facebook has said that it will comply with UNGPs but has much work to do; see Sanjana Hattotuwa, "Making Facebook's New Human Rights Policy Real", Institute for Human Rights and Business 20 April 2021, available: https://www.ihrb.org/focus-areas/information-communication-technology/commentary-making-facebook-new-human-rights-policy-real.

33 UNGP 15 (n 4).

34 OECD Due Diligence Guidance for Responsible Business Conduct (n 9).

35 ISO 31000: 2018 *Risk Management- Guidelines*; see also ISO Guide 73, *Risk Management – Vocabulary*; see also The International Finance Corporation, "Guide to Human Rights Impact Assessment and Management" (2010), available: https://www.unglobalcompact.org/library/25.

36 OECD Due Diligence Guidance for Responsible Business Conduct (n 9), p17.

37 See e.g. ISO, Online Consumer Reviews – Principles and Requirements for their collection, moderation and publication (ISO 20488:2018).

38 OHCHR, B-Tech: "Identifying Human Rights Risks Related to End-Use"; (2020), available: https://www.ohchr.org/Documents/Issues/Business/B-Tech/identifying-human-rights-risks.pdf.

39 Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, (A/74/486), 19 October 2019, para 92, available: https://www.undocs.org/A/74/486 [Accessed 22 July 2021] .

19. The complexity of the risk assessment will vary according to that size of the business, its business model (surveillance capitalism giving rise to distinct risks[40]), its values (including those found in its Community Standards/Terms of Service[41]) and the profile of its users. As noted, companies should be aware of the likely vulnerabilities of groups of users, e.g. the gendered nature of online abuse,[42] or the particular characteristics of children[43].

20. As regards hate speech in particular, social media providers should develop a clear characterisation of online hate – recognising the difficulties around overt and covert expressions of hatred, context (including historical context) and subjectivity – which will feed into the provider's interventions at each of the following four stages:

(1) creation of content (including incentives for certain types of content; tools for creating content (eg emojis, deepfake software and avatars); ease and conditions of access to a platform);

(2) dissemination of content (discovery tools and navigation, including recommender tools, autoplay and virtual assistants);

(3) engagement with content (tools for sharing, responding including likes and votes including users' ability to complain about content); and

(4) deletion of content (moderation and response to complaints and legal actions).

21. These stages may need to be adapted to fit the virtual reality framework but would provide a starting point there; experience from the context of online multiplayer games may be useful here.

22. In setting policies, identifying values and carrying out risk assessments, platforms should be aware of the different levels of hate speech, as well as the 2012 Rabat six-part test[44] for defining incitement to hatred and the application of Article 20 International Covenant on Civil and Political Rights (ICCPR):

(1) the social and political context;
(2) status of the speaker;

40 Zuboff, Shoshana, The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power, (New York: Public Affairs, 2019).

41 The set of rules about expected behaviour on a platform or service, usually against which the platform enforces sanctions.

42 Human Rights Council, Freedom of Expression and Opinion, (A/HRC/44/L.18/Rev.1), 14 July 2020, 8c, available: https://undocs.org/en/A/HRC/44/L.18/Rev.1 [accessed 22 September 2021].

43 ' The risks and opportunities associated with children's engagement in the digital environment change depending on their age and stage of development. They [States parties] should be guided by those considerations whenever they are designing measures to protect children in, or facilitate their access to, that environment. The design of age-appropriate measures should be informed by the best and most up-to-date research available, from a range of disciplines.' General comment No. 25 (2021) on children's rights in relation to the digital environment CRC/C/GC/25 2 March 2021.

44 Report of the United Nations High Commissioner for Human Rights on the expert workshops on the prohibition of incitement to national, racial or religious hatred, 11 January 2013, (UN HCHR A/HRC/22/17/Add.4), para 29, available: https://www.ohchr.org/Documents/Issues/Opinion/SeminarRabat/Rabat_draft_outcome.pdf [accessed 15 November 2021].

(3) intent to incite the audience against a target group;
(4) content and form of the speech;
(5) extent of its dissemination; and
(6) likelihood of harm, including imminence.

23. Social media companies coming to risk assessment for hate speech for the first time should also evaluate its existing risk management practices and processes, practices in relation to human rights impact assessments generally, and data protection/ privacy impact assessments to evaluate any gap or tensions in those practices and processes and ensure that there is appropriate governance[45]. Particular attention should be paid to reliance on techniques driven by machine learning and artificial intelligence and the well known questions around the design and deployment of ML/AI[46].

24. The risk assessment process should be based on data and, where available, research, rather than a hopeful expectation that bad stuff is not happening or, if it is, that it is not the problem of the social media provider. It involves the recognition that the use of technology, including AI, does not in and of itself necessarily ensure human flourishing.[47] It should cover an assessment of actual and potential impacts. This involves gathering data in a systemic manner[48] as to what is happening on the service (e.g. what sorts of user complaints are coming, how are they dealt with), as well as the results of any testing on the product (see below), to understand the nature of the problem, as well as its scale, context and triggers and to acknowledge that information, not bury it.

25. For example, hate speech tends to spike for 24-48 hours after key national or international events such as a terror attack, and then rapidly fall.[49] Systems should be responsive to foreseeable public events (eg major sporting championships), and the due diligence process and mitigations should reflect this. Companies should also bear in mind wider industry experience (e.g. whether certain features – for example live streaming – are particularly risky) and good practice.

26. Where human rights are involved in risk assessment and risk management, their special nature should be recognised, as the OECD due diligence guidance

45 For guidance on human rights-friendly governance procedures, generic to any company type see the UNGPs Interpretative Guide and for technology companies the OHCHR B-Tech project.

46 Council of Europe 'Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems (Adopted by the Committee of Ministers on 8 April 2020), available: https://search.coe.int/cm/pages/result_details.aspx?objectid=09000016809e1154 [Accessed 22 July 2021

47 Ad Hoc Expert Group (AHEG) for the preparation of a draft text of a recommendation on the ethics of artificial intelligence, First Draft of the Recommendation on the Ethics of Artificial Intelligence, 7 September 2020 (SHS/ BIO/AHEG-AI/2020/4 REV.2) (Paris: UNESCO, 2020), para 25, available: https://unesdoc.unesco.org/ark:/48223/pf0000373434 [accessed 26 July 2021].

48 See e.g. Danish Institute for Human Rights in collaboration with the Human Rights Centre at University of Essex 'Guidance on Human Rights Impact Assessment of Digital Activities' (2020), available: https://www.humanrights.dk/publications/human-rights-impact-assessment-digital-activities

49 Matthew Williams and Mishcon de Reya, *Hatred Behind the Screens: A Report on the Rise of Online Hate Speech* (London: Mischcon Academy, 2019), p 24, available: https://hatelab.net/wp-content/uploads/2019/11/Hatred-Behind-the-Screens.pdf [accessed 15 November 2021].

recognises[50]. Companies should respect the need for diversity and inclusion in a risk assessment process so that issues – especially those which particularly affect minorities – are not overlooked or under-valued. This may be particularly relevant when products designed for operation in one state are then deployed in others.

27. A risk assessment should also identify where there is likely to be a conflict of requirements between local laws and internationally recognised human rights to try to identify timely and appropriate responses.[51]

28. Risk assessments should be kept up to date. At the least, risk assessments should be undertaken in relation to a new service before it is deployed, before significant changes or new features are introduced, and if the service is to be deployed (or the provider becomes aware its user base is increasing) in new territories. Risk assessment may link to the measurement of the success of mitigation measures.

Harms

29. Social media companies need to be aware of the harm caused to users by hate speech when assessing risk. Online hate speech gives rise to a number of adverse consequences on victims, some equivalent to the reaction to trauma of 'physical' crimes such as burglary, assault and robbery. Moreover, there is a longevity to the abuse that is distinctive to the online environment.[52] It can also form part of a process leading to hate speech offline and even physical attacks.

> *'The Special Rapporteur shares the concern expressed in one submission that dehumanizing language, often reducing minority groups to animals or insects, normalizes violence against such groups and makes their persecution and eventual elimination acceptable, and that, when committed with a discriminatory or biased intent, these violations become a pathway of demonization and dehumanization that can lead to genocide.'*[53]

30. Note that the harms suffered through hate speech and links forms of aggression may be exacerbated in virtual or augmented reality[54]; given the immersive nature of these technologies, will the experience be more intense than hate speech experienced via text or even video/audio?

31. Some of these harms suffered by victims of hate speech may also be characterised as interference with an individual's human rights, and should be recognised as such

50 'The OECD Guidelines for MNEs state that in the case of human rights, severity is a greater factor than likelihood in considering prioritisation. Thus where prioritisation is necessary enterprises should begin with those human rights impacts that would be most severe, recognising that a delayed response may affect remediability.' P 49 Q5 OECD Due Diligence (n 9).

51 Discussion of UNGP 23 (n 4), "The Corporate Responsibility to Protect Human Rights – an Interpretative Guide (n 8) p 77.

52 Ruth Lewis, Michael Rowe and Clare Wipter, 'Online Abuse of Feminists as an Emerging Form of Violence against Women and Girls' (2017) 57 *Brit. J Criminol* 1462, p. 1478

53 Special Rapporteur de Varennes – 'Thematic report: hate speech, social media and minorities' (n 1), para 44.

54 Lindsay Blackwell, Nicole Ellison, Natasha Elliot-Deflo and Raz Schwarz, 'Harassment in Social Virtual Reality: Challenges for Platform Governance' (2019) 3 Proc. ACM Hum-Comput. Interact No CSCW, Article 100 (Nov 2019), https://doi.org/10.1145/3359202.

in the risk assessment. Examples include Article 17 ICCPR in relation individuals' right to privacy, family, home or correspondence as well as against unlawful attacks on honour and reputation.[55] In its 2018 General Comment, the Human Rights Committee highlighted that:

> *"intentional and negligent homicide, unnecessary or disproportionate use of firearms, infanticide, [honour] killings, lynching, violent hate crimes, blood feuds, ritual killings, death threats, and terrorist attacks"*

32. Are forms of violence that may result in deprivation of life for the purposes of the ICCPR.[56] Moreover, the right to life encapsulates the right to live with dignity.[57] General Comment 37 on the right of peaceful assembly makes the point that Article 21 ICCPR covers online assembly.[58] Rights should moreover be enjoyed without discrimination.[59] Some regional systems have gone further, proposing 'an intersectional and differential focus' which 'takes into consideration the possible aggravation and frequency of human rights violations due to conditions of vulnerability or historic discrimination of persons and collectives'.[60]

33. The fact that harms can constitute human rights violations has consequences for company choices. Normally, the choices are seen as four "T"s:

- Treat - decide on additional controls/mitigation
- Tolerate - Accept the level of risk
- Transfer - pass on the risk to an outside organisation
- Terminate - stop the activity.

34. In the context of human rights 'tolerate' is rarely likely to be acceptable and transfer not at all.

Mitigation process

35. Following the risk assessment, the social media provider must identify appropriate responses to the risks identified. Safety by design (see below) is an important aspect of this. Risk management is dynamic involving assessing the effectiveness

55 See also the developing line of case law in the regional systems, notably that on Article 8 ECHR.

56 Human Rights Committee, General Comment on article 6 of the International Covenant on Civil and Political Rights, on the right to life, (GC 36), 30 October 2018, para 20, available: https://tbinternet.ohchr.org/Treaties/CCPR/Shared%20Documents/1_Global/CCPR_C_GC_36_8785_E.pdf [accessed 21 July 2021].

57 General comment No. 36 (2018), ibid, para 3.

58 Human Rights Committee, General Comment on the right of peaceful assembly (article 21), 17 September 2020, paras 6 and 10, available: https://tbinternet.ohchr.org/_layouts/15/treatybodyexternal/Download.aspx?symbolno=CCPR%2fC%2fGC%2f37&Lang=en [accessed 26 July 2021]. The General Comment does not consider what violence – or non-peaceful conduct – looks like online.

59 Article 2(1) ICCPR; Human Rights Committee, General Comment on Non-discrimination (General Comment 18). The Inter-American Court has considered equality and non-discrimination as *jus cogens*: Advisory Opinion OC-18/03, 17 September 2003, Series A No. 18, para 101.

60 Inter-American Commission on Human Rights, *Business and Human Rights: Inter-American Standards*, (CIDH/REDESCA/INF.1/19), 1 November 2019, para 44, available: https://www.oas.org/en/iachr/reports/pdfs/Business_Human_Rights_Inte_American_Standards.pdf [accessed 21 September 2021].

of responses to risk assessment and, where necessary, adapting the response as appropriate. In determining responses, providers should - in particular - be aware of the privacy and data protection rights of its users. Solving one problem does not automatically justify infringement of other rights.

36. As UNGP 21[61] notes, where there is a risk of severe impact on human rights due to a company's activity public reporting is to be expected. Social media service providers should take proportionate steps to ensure that people who are at risk of human rights impact can learn about potential risks. For large global service providers this might mean publishing in several languages particularly those used by potential or historic victims and in places where victims can easily find such information. Transparency helps inform victims to help them manage their own risk in using services. Transparency allows civil society and public authorities to assist the service provider in mitigating the risks and in managing the difficulties in balancing conflicting rights.

Remediation

37. In terms of assessing risk and remediation, UNGP 24 introduces a three-step hierarchy[62]:

> *'Where it is necessary to prioritize actions to address actual and potential adverse human rights impacts, business enterprises should first seek to prevent and mitigate those that are most severe or where delayed response would make them irremediable.'*

Harm prevention where possible

38. Greater emphasis should be placed on prevention of harm than on remediation, and that the three-stage model in UNGP is a hierarchy, so that rather than engaging in an activity which creates harm and justifying that by some compensatory measures, a company should refrain from the activity[63] This is difficult to apply comprehensively in the context of social media because of its interconnection with freedom of expression; avoiding the issue is not possible. When assessing the risks and the appropriate mitigation strategies, social media service providers should recognise the different dimensions of freedom of expression (the right to speak, the right to be silent, the right to receive and – presumably – the right to ignore or not to receive content). Social media providers may as suggested above be in the position of having to balance conflicting rights, which includes the right to freedom

61 '21. In order to account for how they address their human rights impacts, business enterprises should be prepared to communicate this externally, particularly when concerns are raised by or on behalf of affected stakeholders. Business enterprises whose operations or operating contexts pose risks of severe human rights impacts should report formally on how they address them', *Implementing the United Nations "Protect, Respect and Remedy" Framework* (HR/PUB/11/04 OCHR 2011).

62 UNGP Interpretative Guide (n 8), p82.

63 "When outright elimination of risk is unfeasible, avoidance should be understood as reduction at sources. This introduces a new parameter in human rights due diligence to emphasize that root causes must be identified and addressed." Radu Mares, "Securing human rights through risk-management methods: breakthrough or misalignment?", (2019) 32(3), *Leiden Journal of International Law*, 517, also available here: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3337097.

of expression of all speakers, as well as the other human rights recognised in the UN system. Given that all rights are equal and indivisible[64] it is not possible to see the risk assessment through the prism of freedom of expression alone; other rights would be seen as exceptions to freedom of expression, implicitly downgrading them. The difficulties in reaching this balance, and that fact that there would likely be some infringement of some rights in any position adopted, does not mean that not all mitigating choices are equally appropriate. Companies should place an emphasis on preventing the worst harms (eg incitement to genocide and hate crimes). Moreover, design choices and tools should not have the effect of placing responsibility for protection for users' human rights entirely on those users and this is particularly important where speech is contrary to international or (human rights respecting) national criminal law.

39. The seriousness of the speech should not be assessed on piecemeal basis, removing individual items of speech from their broader context. Social media service providers should be aware of the effect of the constant nature of the intrusion into people's home of online abuse, as well as the risk of a cascade effect[65] of an abundance of online hate (which may affect the likelihood of physical abuse in the real world). One of the first issues a social media service provider should address after risk assessment will be designing systems to be safe.

Safety by Design

40. Safety by design is an approach that recognises the impact of design on behaviour and the role of design in causing harm.[66] In so doing, safety by design aims to prevent diminution of human dignity of minorities by hate speech so far as possible at source[67] by addressing the design of companies' systems and processes and seeking to understand where they contribute or exacerbate the prevalence of hate speech.[68] It seeks to counter the disinhibition effect.[69] More positively, it could include the aspiration of designing inclusively[70] and to keep in mind how design (and business choices) distributes benefits and burdens between different groups of people.[71] Similar to the 'privacy by design' approach, safety by design is preventive,

64 Vienna Declaration (n 31).

65 Williams M. L.and Mishcon de Reya (n 49), p26

66 "The company's actions or decisions– including during design, promotion and marketing - make it more likely that a product or service will be used in ways that cause a harm." UN OHCHR, A B-Tech Foundational Paper, 'Taking Action to Address Human Rights Risks Related to End-Use', September 2020, page 6, available: https://www.ohchr.org/Documents/Issues/Business/B-Tech/identifying-human-rights-risks.pdf.

67 Mares (n 63).

68 This could also reduce the pressure on moderation systems which commentators think are unlikely to be effective without a change in the business model. See eg, Nathalie Maréchal & Ellery Roberts Biddle, "It's Not Just the Content, It's the Business Model: Democracy's Online Speech Challenge", *New America*, 17 March 2020, p. 10.

69 Work on tools and techniques for this is starting in some areas: see e.g. the Prosocial Design Network which lists features and the prosocial consequences they might have and seeks to test them, available: https://www.prosocialdesign.org/.

70 Virtual reality may pose particular challenges – it is more reliant than other technologies on individuals' abilities to control their physical movements, raising questions about how accessible these platforms might be to those with limited mobility.

71 Sacha Constanza-Chock calls for 'design justice': Constanza- Chock *Design Justice: Community-led Practices to Build the Worls We Need* (Cambridge, MA: The MIT Press, 2020), p 23.

not remedial (though design choices and their impact should be kept under review). In terms of 'safety', this can never be absolute, but rather relative to the context – in the same way as we might speak of the safety of someone taking part in dangerous sports – they have an expectation that the equipment they are using is designed to be safe and support them in making dangerous decisions. "Safety" here is the reduction of hate speech-related harms and adverse human rights impacts.

41. In the social media context, specific applications of the principle are relevant to combat hate speech towards minorities. We propose the following principles towards which design choices should be orientated:

maximum safety within the platform to be the default (even if users can choose to change these settings to a riskier option);

safety to be embedded into the design of the product (whilst allowing for updating and improvement, as well as auditing);

permit as much functionality as possible (avoiding unnecessary trade-offs);

safety choices should take into account and be functional for the full life-cycle of the service (and take account of ultimate de-commissioning);

to include transparency and to expect user-centric and rights-respecting choices.

42. Similar to the relationship between privacy by design and privacy enhancing technologies (PETs), there is a link between safety by design and the emerging field of 'safety tech'.[72] By design requires the safety considerations to be built in, not bolted on as an afterthought; a product that is safe by design should itself be rights respecting (including the right to privacy). Where safety tech is supplied by third parties supply chain considerations apply.

Access to the social network and content creation

43. Basic building blocks of safety by design include a sign-up process, tools to create content and Terms of Service (including Community Standards).

Terms of service/community standards

44. Terms of Service constitute the contract between the social media service provider and the user. They are important in communicating the service provider's values; they should therefore reflect those values and in particular respect for human rights[73].

72 United Kingdom Government, "Safety tech providers deliver products and services that enable safer online experiences for citizens" https://www.gov.uk/government/publications/safer-technology-safer-users-the-uk-as-a-world-leader-in-safety-tech. See also an attempt to align global trends in safety tech:Connie Moon Sehat ,"Advancing Digital Safety: A Framework to Align Global Action", World Economic Forum, 29 June 2021. Available here: https://www.weforum.org/whitepapers/advancing-digital-safety-a-framework-to-align-global-action [Accessed 22 July 2021].

73 David Kaye noted that community standards were set without regard to human rights: Report of the Special Rapporteur on Freedom of Opinion and Expression (A/HRC/38/35), 6 April 2018, available: https://undocs.org/en/A/HRC/38/35 [accessed 6 April 2018), p 14 et seq.

They may include Community Standards (though sometimes Terms of Service and Community Standards are used interchangeably) or acceptable use policies, understood as the content and behaviour rules the provider will enforce. Specifically, the Community Standards should make clear the service provider's position on hate speech in each state in which it operates.

45. Many reports of hate speech incidents on social networks reveal deliberately or accidentally deficient Community Standards[74]. In risk management, service providers should look hard at the adequacy of their terms to prevent hate speech. This involves the platform understanding what hate speech is in its various manifestations and levels of severity, and then communicating that to its users. Terms of Service or Community Standards should not bundle different types of hate speech together but should differentiate between them and their various impacts. Insofar as Community Standards are developed based on user reporting, care must be taken to ensure that the rights-based interested of minorities are not overlooked because of issues of volume. Moreover, terms must be understood in sufficient granularity so as to allow for the range of experience for people with the same type of characteristic but who belong to different groups (eg different races, different religions). Terms used to police the boundaries of terms should be checked to ensure that they do not constitute indirect discrimination.

46. When determining its standards, the provider should bear in mind the nature of its service (eg distinction between a general access platform and smaller/niche services) and its user base. Especially where the provider has a broad reach, it should avoid reflecting a narrow world view which 'tend[s] to be typically from the specific sociocultural context of Silicon Valley: racially monochromatic and economically elite'[75]. One option would be to update the Community Standards in consultation with groups who suffer from abuse. There are models for Terms of Service available written with hate speech in mind which could serve as reference points.[76] When making community standards (as well as explanatory guidance) available, social media providers should ensure that appropriate language versions are available, especially for minority groups.

47. Terms of Service should be easily visible before a user signs up to the service, be easy to understand and be available in languages used by the service's users; this is important as part of transparency and processes to hold service providers and, where relevant, users to account. It is the local language versions that should be binding on users, rather than terms of service expressed in the language of the provider's home country. Terms of Service and Community Standards should be kept under review, and revised where appropriate taking into account not just changes in external context but also learning from risk assessments and complaints and moderation processes.

74 'Instagram admits moderation mistake over racist comments' Criddle C. BBC 15 July 2021, available https://www.bbc.co.uk/news/technology-57848106.

75 Report of the Special Rapporteur on Freedom of Opinion and Expression (A/HRC/38/35) (n 22), para 84, citing Ysabel Gerrard and Helen Thornham, "Content moderation: Social media's sexist assemblages", (2020) 22 *New Media and Society*, 1266–1286.

76 Eg https://www.changetheterms.org/terms.

48. Enforcement of Terms of Service does not mean that platforms must actively seek outcriminal content, or monitor generally[77]. Such general monitoring has adverse impacts for all users' freedom of expression and privacy and would be very difficult, if not impossible, to justify.

49. Terms of Service/Community Standards also affect the approach to moderation (discussed further below).

Account creation

50. There has been much concern about anonymous accounts and their role in online abuse and hate speech[78]. Providers should give consideration to Know your Client (KYC) processes, bearing in mind the nature of the platform and its user base. This is not the same as requiring social media service providers to ban anonymous accounts. While the extent to which there is a fundamental right to communicate anonymously is contested[79], it should be recognised that anonymous accounts are important in the protection of minorities, as well as for whistle-blowers and those seeking to hold the powerful to account.

51. Nonetheless, social media service providers should consider the risk of people abusing anonymity to direct hate speech towards minorities and take steps to mitigate that risk, whether in terms of account verification, or through other interventions (e.g. enhanced user self-protection tools that could, say, block unverified accounts; or effective reporting mechanisms). Service providers should assess the risk of harm arising through hate speech from fake identities (eg those used for catfishing[80] or sock puppet accounts[81]); whether multiple accounts per person are permitted (and in what circumstances); and whether bots should have accounts.[82] Providers should take proportionate steps to address these risks; this should go beyond a mere statement in terms of service that such behaviour is prohibited. Service providers should consider whether those who have been banned (for a period) from the service should be prevented from circumventing that ban. The provider could consider whether more friction could be introduced into the process, eg a cooling off period after sign up.

77 Note there is a difference between monitoring (eg via an upload filter) which looks for specific content (eg on the basis of hashes or watermarks) and that which searches communications generally. Within the EU, the former is acceptable: Case C-18/18 *Eva Glawischnig-Piesczek v Facebook Ireland Limited,* judgment 3 October 2019, available: https://curia.europa.eu/juris/document/document.jsf?text=&docid=218621&pageIndex=0&doclang=en&mode=req&dir=&occ=first&part=1&cid=2137970 [accessed 22 September 2021].

78 UK Parliament debate: Online Anonymity and Anonymous Abuse Volume 691: debated on Wednesday 24 March 2021 Available at https://hansard.parliament.uk/commons/2021-03-24/debates/378D3CBD-E4C6-4138-ABA6-2783D130B23C/OnlineAnonymityAndAnonymousAbuse [Accessed 26 July 2021].

79 Barendt, Eric *Anonymous Speech: Law Literature, Politics* (Oxford: Hart Publishing, 2016).

80 Where a person creates a fake identity to take advantage of another user.

81 An online identity used for deception, often for the purpose of talking about or to themselves while pretending to be another person; the term is now used more broadly to include those manipulating public opinion, to circumvent restrictions, such as viewing a social media account that they are blocked from, suspension or an outright ban from a website. They are different from pseudonyms. See eg Farkas et al 'Cloaked Facebook Pages: exploring Fake Islamist Propaganda in Social Media' (2018) 20(5) *New Media and Society* 1850.

82 Julia Hass, Freedom of the Media and Artificial Intelligence, OSCE 16 November 2020, p. 4, available: https://www.osce.org/files/f/documents/4/5/472488.pdf [accessed 26 July 2021]; Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini and Filippo Menczer 'The Spread of Low-Credibility Content by Social Bots' (2018) 9 *Nature Communications* 4787, available: https://www.nature.com/articles/s41467-018-06930-7.

52. Service providers should also seek to understand any risks created by networks of accounts (eg coordination and amplification of posts).[83] The concern is the way such networks increase not just the spread but also the speed of dissemination of hate speech, including across different platforms.[84] In this context, service providers could seek to understand who are the direct and indirect instigators and beneficiaries of such hate speech. Service providers could also seek to understand who is operationalising those messages and how, returning us to questions of bots, sock puppets networks and false identities. Some individuals or small groups of individuals might be significant nodes in networks of hate speech that are amplified within the service[85]. Companies should have a transparent process for managing such individuals, carrying out the necessary balancing of human rights. In an interconnected world, service providers might factor into their risk assessment whether and how the highest risk individuals spread hate speech on other services.

Content creation – service design that might increase hate speech

53. Each service is designed to allow and incentivise a person to create content in a different way. How content creation is designed can affect the risks of hate speech being created and disseminated. Features such as metrics (and it is possible to manipulate social media through the purchase of fake engagement[86]) as well as financial incentives based on popularity should be considered in the light of any skew they introduce into content. Outrage and negative emotions (including hate speech) seemingly drive engagement (as clickbait headlines show)[87], and there is a risk of a cycles of ever increasingly outrageous content to drive likes and upvotes.[88] In some cases, making highly harmful content can even be profitable for the social

83 This can be especially problematic when the corordination is the responsibility of a State, and this seems to be a particular problem as, for example, Facebook notes: Meta, *Detailed Report – October 2021 Coordinated Inauthentic Behavior Report*, October 2021, available https://about.fb.com/wp-content/uploads/2021/11/October-2021-CIB-Report.pdf, p. 5. See also e.g Marcel Schliebs, Hannah Bailey, Jonathan Bright and Philip N. Howard, *China's Inauthentic UK Twitter Diplomacy A Coordinated Network Amplifying PRC Diplomats* (Dem.Tech Working Paper 2021.2), 11 May 2021, available: https://demtech.oii.ox.ac.uk/research/posts/chinas-inauthentic-uk-twitter-diplomacy-a-coordinated-network-amplifying-prc-diplomats/ [accessed 21 July 2021].

84 See eg T. Shephard et al 'Histories of Hating' (2015) 1(2) *Social Media and Society* 1, DOI:10.1177/2056305115603997.

85 Renee DiResta et al., *New Knowledge, The Tactics & Tropes of the Internet Research Agency* 42 (2019); Brian Fishman, *Crossroads: Counter-Terrorism and the Internet*, (2019) 2 *Tex. Nat'l Sec. Rev.* 82, 86–87. See by analogy the small network of individuals allegedly responsible for a substantial proportion of anti-vax content 'Just twelve anti-vaxxers are responsible for almost two-thirds of anti-vaccine content circulating on social media platforms. This new analysis of content posted or shared to social media over 812,000 times between February and March uncovers how a tiny group of determined anti-vaxxers is responsible for a tidal wave of disinformation' '*The Disinformation Dozen Why platforms must act on twelve leading online anti-vaxxers*' Center for Countering Digital Hate Ltd (UK) https://www.counterhate.com/disinformationdozen [Accessed 22 July 2021].

86 See eg S. Bay and R Fredheim, *Falling Behind: How Social Media Companies are Failing to Combat Inauthentic Behaviour Online* (NATO Strategic Communications Centre of Excellence, 2019).

87 As recognised in the Report of the Secretary General, State of global peace and security in line with the central mandates contained in the Charter of the United Nations (A/74/786), 6 April 2020, para 43, available: https://undocs.org/en/A/74/786 [accessed 22 September 2021. The problems are acknowledged in Facebook's Civil Rights Audit (July 2020).

88 W. J. Brady et al 'How Social Learnings Amplifies Moral Outrage Expression in Online Social Networks' (2021) (paper under review, available: https://psyarxiv.com/gf7t5/); Soroush Vosoughi, Deb Roy and Sinan Aral, "The Spread of true and false news online" (2018) 6380 *Science* 1146-51, DOI: 10.1126/science.aap9559;.

media platform[89] or some 'content creators' who spread hate to attract likes to earn a living. Addressing some of the concerns around content curation and recommender tools may help, but services providers should seek to understand if there are other features of the platform that might be exploited. [90]

54. The operation of social media platforms has led to the emergence of highly popular new communications media such as hashtags, emojis, photo-filters[91] (including the making available of tools to create filters[92]), deepfakes and the use of GPT3[93] and extending beyond augmented reality to virtual reality. Service providers have often adopted these and encourage their use in content creation to the extent that they become a major feature of some services. Some of these can be used to benefit minorities and celebrate diversity.[94] Sadly these media can be abused to become hate speech and worryingly AI systems trained to identify text-based hate are (perhaps unsurprisingly) much less effective at identifying emojis.[95] Service providers should include such tools and media in their risk assessment and mitigation plans, taking into account the impact they may have on the effectiveness of safety tools and other risk mitigation measures in place, and also consider supply chain issues where content creation and curation tools are provided by third parties.

55. These issues are a starting point. We recommend that more work is undertaken to understand how features can cause problems with a view potentially to expanding this list.[96] A consideration of content creation then leads naturally to how that content is discovered and navigated by users.

89 "Despite promises to keep users safe, we show how Big Tech itself makes up to $1 billion a year in advertising and other revenues from this industry, which threatens the effectiveness of a future Coronavirus vaccine.": Centre for Countering Digital Hate, The Anti-Vaxx Industry: How Big Tech powers and profits from anti-vaccine misinformation, (2020), available: https://www.counterhate.com/anti-vaxx-industry [accessed 15 November 2021].

90 DRFLab, "#InfluenceForSale: Venezuela's Twitter Propaganda Mill", *Medium* 4 February 2019, available: https://medium.com/dfrlab/influenceforsale-venezuelas-twitter-propaganda-mill-cd20ee4b33d8 [accessed 21 July 2021].

91 Morgan Jerkins, 'The Quiet Racism of Instagram Filters' Racked, 7 July 2015, available: https://www.racked.com/2015/7/7/8906343/instagram-racism [accessed 4 November 2021]; Lauren Michele Jackson 'We need to talk about Digital Blackface in Reaction GIFs' (2017) *Teen Vogue* 2 August 2017, available: https://www.teenvogue.com/story/digital-blackface-reaction-gifs [accessed 4 November 2021]; Sarah Lee 'Instagram filters: 'Our skin is for life, not for likes' BBC News, 19 October 2020, available: https://www.bbc.co.uk/news/uk-england-london-54360146 [accessed 4 November 2021].

92 Matte Wille 'Facebook banned blackface. Judging by Instagram filters, you'd never know' 12 Oct 2020, available: https://www.inputmag.com/culture/facebook-banned-blackface-judging-by-instagram-filters-youd-never-know [accessed 4 November 2021].

93 See eg B. Buchanan et al Truth, Lies, and Automation: How Language Models could Change Disinformation, (Center for Security and Emerging Technology, 2021), doi: 10.51593/2021CA003.

94 See eg https://perma.cc/4UXB-KKQM [accessed 4 November 2021]

95 'AI's coming home: How Artificial Intelligence Can Help Tackle Racist Emoji in Football' Hannah Kirk Oxford Internet Institute Blog 16 July 2021 https://www.oii.ox.ac.uk/blog/ais-coming-home-how-artificial-intelligence-can-help-tackle-racist-emoji-in-football/ [Accessed 22 July 2021].

96 A model could perhaps be the survey work undertaken by the OECD on the approach to terrorist and violent extremist content: Current approaches to terrorist and violent extremist content among the global top 50 online content-sharing services OECD August 2020 No.296, available https://www.oecd-ilibrary.org/science-and-technology/current-approaches-to-terrorist-and-violent-extremist-content-among-the-global-top-50-online-content-sharing-services_68058b95-en.

Discovery and Navigation

Presentation of content to users

56. It is common in social networks to use software to select, rank and present or recommend items of content to users and to suggest text while typing.[97] Often this software contains machine learning or 'artificial intelligence'. Machine learning derives its capability from processing large data sets to inform its actions. Minorities will often not be well represented in large datasets used to feed machine learning both from being a statistical minority and compounded in some cases by the minority groups not being sufficiently online to generate a representative amount of data. Also the people who write the machine learning software may well be unaware of or unfamiliar with discrimination against minorities at all or in distant overseas markets where the software is used/applied[98].

57. Many such systems are often described as 'black box' in that their internal workings are not readily visible. The problems that arise from the use of ML/AI are not inevitable (or at least not all); the decision-making processes around their development and deployment must be scrutinised.[99] Even 'black box' systems have outputs, which can be tested. At the statistical scale at which many social networks operate, issues of bias should be discernible. Testing (see below) should take into account how the tool could be used; the experience of Microsoft's chatbot Tay[100] that taught itself to be racist is a warning example.

58. There have been concerns that the effect of the recommender algorithms, especially in conjunction with auto play can prioritise extreme speech, and therefore has a role in spreading hate speech.[101] Service providers should consider what values are embedded in their recommender tools; ideally recommender tools should prioritise credible and authoritative information[102], rather than for prioritising outrage. Additionally, providers should ensure that their recommender features are auditable,[103] including considering and documenting the questions of what was

97 T. Gillespie *Custodians of the Internet* (New Haven/London: Yale University Press, 2018), p. 7

98 Frederik Zuiderveen, "Discrimination, Artificial Intelligence and algorithmic decision making", Borgesius Professor of Law, Institute for Computing and Information Sciences (iCIS), Radboud University Nijmegen, and Researcher at the Institute for Information Law, University of Amsterdam(the Netherlands) Study for Directorate General of Democracy, Council of Europe 2018

99 Committee of Experts on Internet Intermediaries, *Algorithms and Human Rights: Study on the human rights dimensions of automated data processing techniques and possible regulatory implications*, (MSI-NET) Council of Europe study DGI(2017)12, p.8.

100 Oscar Schwartz, "'In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation" Institute for Electrical Engineers 25 November 2019, available: https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation.

101 E. Hussein et al, 'Measuring misinformation in video search platforms: An audit study on YouTube' (2020) Proceedings of the ACM on Human-Computer Interaction, 4(CSCW1), Article 48. doi 10.1145/3392854; S. Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York: NYU Press, 2018)

102 Eg Twitter is rolling out a 'pre-bunking' strategy to try to counter climate change disinformation: https://blog.twitter.com/en_us/topics/company/2021/-cop26-is-happening-on-twitter [accessed 4 November 2021]. Facebook has added information labels to some posts directing users to a centralised source of accurate information on climate change.

103 The issues of explainability have been discussed following the GDPR's inclusion of a right to an explanation; See eg Margot E Kaminski 'The Right to Explanation, Explained' (2019) 34 *Berkley Technology Law Journal* 189, DOI: https://doi.org/10.15779/Z38TD9N83H. Some consider interpretability a better approach: Cynthia Rudin, 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead '(2019) *Nature Machine Intelligence* 206-15.

considered when setting up the features and what the operation of the features show.[104] Any such assessment should consider the differential experience of different groups of user.[105] In this, providers should pay special regard to special guidance[106] on algorithmic accountability[107] and auditing.

59. Autoplay operates to push content at users without those users having chosen to engage with content, affecting a person's freedom to choose the content with which to engage. There has been concern that this, combined with the operation of the recommender machine selecting the content to be pushed, has resulted in the prioritisation of hate speech (amongst other types of illegal and/or unwanted content), concerns which seem to have been validated by the leaked Facebook papers. Note that the likely harm caused by this sort of preference could be worse when dealing with a platform on which there are a large number of users; the platforms should take their size – as well as prevalence of content types – into account too.

60. If autoplays are to be included, providers should consider whether other options have fewer adverse impacts; for example, autoplay only operating with user-selected play lists. Transparency to the user about the basis on which a recommendation has operated, with more specificity than generalities such as 'based on your earlier viewing' can help identify problems that might arise with hate speech and may assist with dispersing some of the opaqueness around how the user's information environment is shaped[108]. Note that in addition to the concerns about filter bubbles arising from personalisation[109] (the existence of which are subject to some debate, or the likelihood of which may vary from platform to platform based on design choices[110]),

> *[f]ine grained, sub-conscious and personalised levels of algorithmic persuasion may have significant effects on the cognitive autonomy of individuals and their right to form opinions and take independent decisions.*[111]

104 See eg Joschua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson and Harlan Yu, 'Accountable Algorithms' (2017) 165 *University of Pennsylvania Law Review* 633, available: https://scholarship.law.upenn.edu/cgi/viewcontent.cgi?article=9570&context=penn_law_review [accessed 26 July 2021].

105 Megan McCluskey, 'Why Some People See More Disturbing Content on Facebook than Others, According to Leaked Documents' Time 3 November 2021, available: https://time.com/6111310/facebook-papers-disturbing-content/ [accessed 4 November 2021].

106 Eg Ada Lovelace and Open Government Partnership, *Algorithmic Accountability for the Public Sector Report*, 24 August 2021, available: https://ainowinstitute.org/pages/algorithmic-accountability-for-the-public-sector-report.html [accessed 4 Nov 2021]. On discrimination more generally see Nicholas Schmidt and Bryce Stephens, *An Introduction to Artificial Intelligence and Solutions to the Problems of Algorithmic Discrimination*, 8 November 2019, arXiv:1911.05755 [cs.CY], available: https://arxiv.org/abs/1911.05755 [accessed 4 November 2021].

107 Note academic criticism of this concept: eg J Kemper and D Kolkman 'Transparent to whom? No algorithmic accountability without a critical audience' (2019) 22:14 I*nformation Communications and Society* 2081, available: https://doi.org/10.1080/1369118X.2018.1477967; R Binns, 'Algorithmic Accountability and Public Reason' (2018) 31 *Phiolos. Technol.* 543, available: doi.org/10.1007/s13347-017-0263-5; M Wieringa 'What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability' (2020) *FAT 20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* 1, available: doi.org/10.1145/3351095.3372833.

108 The UN Special Rapporteur noted the problem of opaqueness: Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (A/73/348), para 10, 26 October 2018, available: https://undocs.org/en/A/73/348 [accessed 26 July 2021].

109 E. Pariser, *The Filter Bubble: What the Internet is Hiding From You* (London: Penguin, 2012) C. R. Sunstein, *Echo Chambers: Bush v. Gore, Impeachment, and Beyond*, (Princeton University Press, 2001)

110 M. Cinelli, G De Franccisci Morales, A Galleazzi et al 'The Echo Chamber Effect on Social Media' (2021) 118(9) PNAS 2023301118, available: doi.org/10.1073/pnas2023301118.

111 Council of Europe, Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic processes (Decl(13/02/2019)1), 13 February 2019, para 9, available: https://search.coe.int/cm/pages/result_details.aspx?objectid=090000168092dd4b [accessed 26 July 2021].

61. Moreover, personalisation breaks down the assumption of a common market place of ideas in which ideas can be challenged and counter narratives asserted. This is a threat to the one justification for free speech – that ideas can be tested through contestation.

62. Autocompletes are a particular subset of the use of automated discovery tools and they operate to define a user's text entry or search term and thus the material that comes to that users attention. Some autocomplete functions suggest racist or abusive searches[112], potentially contributing to the promotion of that message as well as being harmful to those in the relevant group. Providers should consider the adverse impacts of the use of such tools, as well as the nature and extent of any compensatory moderation/removal policies in this context. Reporting features for problematic autocompletes should be clearly visible and easy to use. Where problems arise, providers should verify that the issue actually is solved and that solutions work.

63. Some of these problems can be avoided if service providers are clear about their values and ensure that their recommendation and curation features embody those values, documenting how this has been done – what issues they have sought to address and how.

Advertising

64. It should not be possible for a purveyor of hate speech to buy their way around measures set up to combat hate speech through purchasing an advertisement.[113] Equally advertisers should be able to choose that their adverts are not positioned alongside hate speech thereby in some way fund it.[114] Targeted advertising is the process whereby adverts are sent to particular users only, based on the provider's views of that person's characteristics. Advertisers may choose the characteristics on the basis of which the adverts are delivered. Some characteristics offered to advertisers to buy against are generated automatically by the service provider's software. Concern about the possibility for discrimination has already arisen; there is a concern that this sort of feature may be used either to target minorities or to develop support for those seeking to discriminate or oppress minorities. This form of targeting may run into the same sorts of concerns about AI manipulation and concerns about the operation of the market place of ideas and the search for truth. Conversely, systems designed to protect against discrimination should ensure that they do not inadvertently prevent content creators from minoritised groups from earning ad

112 Seth Stephens-Davidowitz, 'Hidden hate: What Google searches tell us about antisemitism today' (Antisemitism Policy Trust and Community Security Trust, 2019) available at https://archive.jpr.org.uk/object-uk508 [Accessed 26 July 2021]

113 Note also the issue of buying fake engagement, discussed earlier.

114 "Facebook, YouTube and Twitter have agreed a deal with major advertisers on how they define harmful content. Brands will also have better tools to control where their ads appear. It follows an advertising boycott of Facebook earlier this year, involving more than 1,000 companies." BBC Reporter, 'Advertisers strike social media deal over hate speech', 23 September 2020, available: https://www.bbc.co.uk/news/technology-54266534 See also the work of the Global Alliance on Responsible Media for improving online safety for advertisers. https://wfanet.org/leadership/garm/about-garm

revenue.[115] Service providers should include their advertising mechanisms in a hate speech risk assessment and consider how oversight – particularly for ML/AI enabled features, is to be implemented. Providers should also consider the usefulness of maintaining an advert library as part of the provider's oversight arrangements.

65. When hate speech evades the risk assessment and mitigation process described above user defence tools should be considered.

User Response, User Tools

66. In general terms, the providers should ensure that user tools to engage with the content of other users cannot easily be weaponised, used in a discriminatory fashion, and where they need counter-measures that these measures are effective, accessible and easy to use.[116]

User Self-care Tools

67. As part of their risk management, social media companies should provide tools for users that can be used if systemic risk mitigation fails.

68. User tools are mechanisms that allow users to provide 'digital self-care'. They form part of users' ability to exercise some control over their online environment. The platform makes design choices about whether to provide these tools and how easy they are to find and use (including providing instructions and examples in relevant languages). Given the tendency of users not to change the original settings, providers should give strong consideration to enabling safety settings as default.

69. For members of minority groups, such tools can be an important part of ensuring that human rights impacts do not occur or are limited if other systemic measures have failed. It is not desirable to transfer a burden to the victim, but it is a prudent backstop. User tools seem appropriate for lesser examples of hate speech; it would seem disproportionate to expect users to deal with criminal hate speech by using blocking tools alone.

70. Muting and blocking tools might give rise to concerns about the rights of the speaker and 'filter bubbles'; would counter-narratives be suppressed? The right to freedom of expression limits the ability of states to intervene in communication between willing speaker and willing listener but does not give a speaker the right to force someone to listen to that speaker. Nonetheless, the right to receive presumably also implies the right not to receive, though like the expressive right, it is not unlimited. In implementing user tools, the platform should consider the impact on counter narratives, as these tools would be available to the perpetrators of hate speech as

115 CHEQ, *Brand Safety's Technological Challenge: How Keyword Blacklists are Killing Reach and Monetization*, September 2019, available: https://info.cheq.ai/hubfs/Research/Brand_Safety_Blocklist_Report.pdf [accessed 4 November 2021]

116 For an example of this sort of problem see J. Nathan Matias: https://twitter.com/natematias/status/1296219943743168518 [accessed 4 November 2021]; Anna Kramer, 'What Tracy Chou learned about online harassment while building an app to solve it' Protocol 26 January 2021, available: https://www.protocol.com/harassment-block-party-app [accessed 4 November 2021].

well as victims of it. In designing tools providers should consider (and document) how they determined an appropriate balance. Some of these tools may also operate to protect a user's privacy, and this should be borne in mind when carrying out risk assessment and mitigation proposals.

Complaints processes

71. Linked to Community Standards and moderation processes and complaints mechanisms. Complaints processes provide vital early warning of hate speech problems. The UNHCR B-Tech guidance[117] recognises that:

> *Company-based grievance mechanisms have particular value as an "early warning system" with respect to the human rights implications of a company's business activities. They also provide a source of information which can be used to analyse trends and the effectiveness of corporate responses to human rights risks.*

72. The adequacy of complaints processes should be part of hate speech risk assessment. By implication, companies should devote resources to the problem at appropriate scale and for each linguistic group of users.

73. The provider should also ensure that the design of complaints mechanisms is user-centric: that is, visible, easy to use and age and language appropriate. Complaints processes should not just be limited to complaints about individual items of content. They should allow for complaints about a series or pattern of communications as well as to features of the services itself (eg the way the recommender algorithm works, or other 'dark patterns' and nudges, or tools for creation).

74. As noted earlier, UNGPs 29 and 31 emphasise the need for grievance processes and these obligations reflect that, dealing with issues of accessibility, predictability, equitability, transparency. It is not just users of a particular platform that might be affected by hate speech on it and that non users too should have some rights to complain. Reasoned decisions are important as a form of continuous learning about the nature of the problem.

75. Reasoned decisions' primary addressees are those affected by the decision and the decision should be formulated with that in mind; best practice suggests that the decision should take express notice of the rights involved. Any such decision should give clear instructions for rights of challenge and remind users of any rights before the courts under domestic law. Grievance mechanisms, although useful, do not replace the role of independent courts.

76. The Interpretative Guidance to the UNGPs suggest that companies when they find themselves at fault in the grievance process should consider the views of victims

117 OHCHR blog post - 'This blog post sets out key expectations for technology companies under the UN Guiding Principles on Business and Human Rights 'January 2021' from https://www.ohchr.org/Documents/Issues/Business/B-Tech/B-Tech-Blog-USCapitolResponse.pdf and also B-Tech Foundational Paper 'Designing and implementing effective company-based grievance mechanisms' https://www.ohchr.org/Documents/Issues/Business/B-Tech/access-to-remedy-company-based-grievance-mechanisms.pdf [accessed 22 July 2021].

when considering remediation:

> *This may be an apology, provisions to ensure the harm cannot recur, compensation (financial or other) for the harm, cessation of a particular activity or relationship, or some other form of remedy agreed by the parties.*[118]

Speed of transmission

77. Many providers aim to ensure communication is as frictionless as possible, which means that people can share content even without reading, certainly not considering it (and similar points may be made about like buttons and similar features), supporting the virality of certain sorts of content. This is potentially problematic given the bias towards content expressing negative emotion. Social media service providers should therefore consider the constitutive role of these features in the spread of hate speech, particularly hate speech of the type that is contrary to Article 20 ICCPR.

78. When facing a human rights crisis, some service providers have quickly reached for tools to limit velocity on the platform, recognising that this to some extent mitigates the problem of harmful messaging.[119] Frictionless communication may also run the risk that people engage without thinking and allow for rapid propagation of hate speech with severe impacts on human rights. Some prompts to speak have been seen as exerting pressure on users (e.g. the impact of Streaks on Snapchat). It is an open question about whether these undermine the autonomy of individuals as regards their freedom of speech.

Moderation

79. Groups representing victims of hate crime, including trusted flaggers, reported to us that where harm arose from hate speech, adequate moderation processes were often either not in place or were not sufficiently resourced proportionate to the risk. This suggests that in their approach to moderation social media companies have taken a risk management decision in favour of shareholders rather than minorities.

80. Here, we assume trusted flaggers to mean an individual or organisation that has particular expertise and competence in detecting, identifying and providing notification of a specific category of illegal content and may include hotlines (eg Internet Watch Foundation in relation to child abuse). Service providers should consider the need for, or refer to, appropriate, accreditation or verification process that would independently evaluate the quality of the notices provided by trusted flaggers. They should make information about "trusted flaggers" available on their websites.

81. We note the point made by the Special Rapporteur on Minorities that:

118 Interpretative Guide (n 8) Q64 .

119 'WhatsApp has said it will limit how many times messages can be forwarded in India, to curb the spread of false information on its platform. The announcement comes after a spate of mob lynchings were linked to messages that circulated on WhatsApp groups.' BBC Reporter. 'India lynchings: WhatsApp sets new rules after mob killings' 20 July 2018, available: https://www.bbc.co.uk/news/world-asia-india-44897714.

> *In order to improve mechanisms and content vetting policies for the handling of hateful content, and to ensure incorporation of the concerns of the main targets of hate speech in social media, the Special Rapporteur urges that minorities, as the most targeted and vulnerable groups, be represented in advisory and other relevant boards.*[120].

82. This exhortation is consistent with the UNGPs suggesting that companies involve victim groups in preventing human rights abuses. There is concern that currently minority groups are not adequately involved in the determination of what counts as hate speech.[121]

83. As part of their hate speech risk assessment, companies should assess what form of moderation is appropriate, whether it is in house, whether it relies on external volunteers (including trusted flaggers), or whether some form of automation should be used.[122] Platforms should also ensure that moderation teams are sufficiently resourced in relation to each territory that the services serves; language issues should also be considered.[123] As noted, moderation is more difficult with regard to emojis; similar questions arise with regard to augmented reality tools.

AI and Moderation

84. There is already a significant literature on AI and content moderation.[124] This commentary does not seek to repeat that but to emphasise that use of such tools is no silver bullet. Use must be carefully assessed, bearing in mind known difficulties about AI systems generally relating to accuracy and bias but also those specific to content moderation: that the expression of marginalised communities may be improperly labelled as hate speech (including where slurs are reclaimed) as well as the usual difficulties in assessing context. AI tools may not operate equally well in all language contexts, and may prioritise text over images. Context and humour are difficult to deal with through automation. Particular attention should be paid to the question of whether such systems adequately recognises intersectionality, and whether AI works equally well for all forms of content/all types of users. As with the role of AI in content amplification, platforms should consider the possibility of allowing users not to be subject to AI moderation. In any event, they should inform users if and how AI is used in easy to understand terms, and always have a human in the loop. We have noted the importance of reasoned decisions; that reasoning is relevant here too.

120 SR De Varennes: Thematic Report (n 1) para 98.

121 Eugenia Siapera 'AI Content Moderation, Racism and (de)Coloniality' (2021) *International Journal of Bullying Prevention,* https://doi.org/10.1007/s42380-021-00105-7.

122 Robyn Caplan has categorised types of moderation: R Caplan, "Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches", 14 November 2018, Data and Society Report, available: https://datasociety.net/library/content-or-context-moderation/.

123 Stephanie Valencia 'Misinformation online is bad in English. But it's far worse in Spanish' *Washington Post* 28 October 2021.

124 See for example OSCE Guidelines, (n82).

Impact of moderation

85. The service provider should ensure that the moderation response adopted is proportionate to the harm/intensity of hate speech and that they provide a reasoned decision that explains this linkage. The decision should bear in mind the importance of freedom of expression to democracy. Consideration of democracy should not, however, mean an automatic exemption for politicians who express hate speech; in principle the same standards should be applied to all (in the interests of equal enjoyment of rights). The clear articulation of the hate speech policy is fundamental to facilitate such deliberation and explanation.

Take down considerations

86. Social media service providers' approach to take down as a tool to combat hate speech should also be part of the risk assessment. Search providers should consider whether take down (or account suspension/removal) is a proportionate response. In terms of take down times, a provider should consider what is appropriate, and even what is the appropriate measure of responsive take down times – does it depend on the time elapsed, or the number of impressions? This might depend on the nature of the content and the types of person harmed.

87. In the case of something like the Christchurch massacre it may be that impressions are an appropriate measure because virality of that content was the concern in that instance; time taken to respond might be a more appropriate measure when looking at content directly addressed to individuals.

88. Linked to this is the more difficult question of content stay-downs, given the strict limitations imposed on filtering required by governments.[125] However, it seems clear that the fact that stay downs concern content that have already been found to be in violation of the rules and where take down was legitimate perhaps shift the assessment from a freedom of speech perspective; there may still be privacy issues.

89. Note also that it is possible that providers may choose to remove content without waiting for a complaint within the Terms of Service/Community Standards.

Safety Testing

90. Safety testing should be at the heart of due diligence and risk assessment, a position reflected by UNGP 15. Testing is particularly relevant in an approach that focuses on very complex software systems. For over 150 years, scientific testing[126] of company processes has been intrinsic to protecting people from harm – both workers,

125 Emma J. Llanso, 'No amount of "AI" in content moderation will solve filtering's prior-restraint problem' (2020) *Big Data and Society* 1-6, available: https://journals.sagepub.com/doi/pdf/10.1177/2053951720920686; Emma Llanso, Joris van Hoboken, Paddy Leerssen and Jaron Harambam, *Artificial Intelligence, Content Moderation, and Freedom of Expression* (Transatlantic Working Group, 2020), available: https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf.

126 Peter W.I. Bartrip, "The State and the Steam-Boiler in Nineteenth Century Britain", (1980) 25(1) *International Review of Social History* 77-105.

customers and people who might be harmed by, but are not involved in the company or its products. External testing standards work best when they are transparent and, for the most hazardous services are carried out by independent people. In some industries there are multinational agreements on testing procedures and standards to protect the public. We understand that social media service providers carry out extensive testing of product features to meet their commercial goals[127] but testing for safety seems less exhaustive or are ignored at a senior level in the company[128].

91. Social media providers should carry out testing and examination of their services, preferably prior to deployment, that enables them at a minimum to:

(a) understand whether the measures they have put in place are working to prevent, or appropriately mitigate, the incidence of hate speech to minorities; and

(b) detect whether new forms of hate speech to minority groups have appeared. The service provider should also test for/measure whether the measures in place to protect minorities have unduly restricted other rights.

92. Companies should test for harm to each of the minority groups it identified in its risk assessment. Confidence in the service provider would be enhanced if it published the results of such testing in a timely manner as well as allowing external review.

93. Testing should not be carried out solely against a standard, but also involve exploratory, qualitative investigation to assess exactly how a new feature could be used to convey hate speech at each stage of the four stage model set out in above (and include consideration of how decentralised models might be used). Such testing will work best if it involves people who have experience of being a victim of hate speech.[129]

Supply Chain Issues

94. It is increasingly common for social media service providers to contract out parts of their business function, which could have an impact upon the human rights of minorities.

95. UNGPs 13 and 18 which concern supply chain issues are instructive here. Social media service providers should not seek to avoid responsibility for human rights

127 'Facebook engineers and data scientists posted the results of a series of experiments called "P(Bad for the World)." The company had surveyed users about whether certain posts they had seen were "good for the world" or "bad for the world." They found that high-reach posts — posts seen by many users — were more likely to be considered "bad for the world," a finding that some employees said alarmed them.' Kevin Roose, Mike Isaac and Sheera Frenkel, 'Facebook Struggles to Balance Civility and Growth', New York Times, 24 November 2020, available: https://www.nytimes.com/2020/11/24/technology/facebook-election-misinformation.html.

128 'I saw that Facebook repeatedly encountered conflicts between its own profits and our safety. Facebook consistently resolved those conflicts in favor of its own profits. The result has been a system that amplifies division, extremism, and polarization — and undermining societies around the world' Haugen F. Statement United States Senate Committee on Commerce, Science and Transportation Sub-Committee on Consumer Protection, Product Safety, and Data Security October 4, 2021.

129 'People on Reddit and Twitter started taking screenshots of the KFC emoji …. The connotation is extremely racist, and plays into antiquated stereotypes. Seeing the two emotes being used together gave people pause for concern. Do, like, no black people work at Twitch? What else do I even say to that? 'Abuse of KFC emote on Twitch leads to more conversations about toxic chat culture' Julia Alexander Mar 26, 2018 Polygon https://www.polygon.com/2018/3/26/17163582/kfc-emote-twitch-trihex-forsen-trihard-xqc [Accessed 22 July 2021].

impacts through outsourcing or ignore the human rights and harms risks arising from it – even if, as UNGP 13 states, those providers 'have not contributed to those impacts'. Risk assessments therefore should include an assessment arising from business relationships.

96. The impact could be felt by people who are customers or people who work for the supplier such as a moderators with poor working conditions – and little time to make decisions[130]. This trend could continue as application programming interfaces allow componentisation of a software service, sometimes in response to regulatory pressure.[131] Large social networks will be able to apply leverage (as the UNGPs suggest) to ensure that sub-contractors or suppliers follow international human rights norms. Smaller social networks might have to consider whether contracting out some components is worth the human rights risk for their customers.

97. Note also that some features may not be derived from a formal business relationship but be through third part independent software such as services that allow a user to post to multiple social networks. Social media providers should also consider the risks of harms arising from hate speech arising from such software. Likewise, should service providers use a decentralised model, consideration should be given as to how that model might function and what safety features might operate and how.

Victim Support and Remediation

98. Social media service providers should consult victims and victim-representative groups in a respectful and sensitive manner to design remedies for people who have been harmed by exposure to hate speech.

99. Victim support can be of help in mitigating the harm suffered by victims of hate speech, at least that at the lower end of the hate speech scale, and go some way towards providing rehabilitation[132] and potentially form a (small) part of the "remedy ecosystem"[133]. The provider can facilitate users finding this support, as not all such organisations are known about or visible. Victim support is not a substitute for nor an alternate to stopping hate speech at source (as set out above).

130 S. T. Roberts *Behind the Screen: Content Moderation in the shadows of social media* (Yale University Press, 2019); N Hopkins 'Revealed: Facebook's internal rulebook on sex, terror-ism and violence' in The Guardian, 21 May 2017, available: https://www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence [accessed 4 November 2021].

131 "Evidence suggests that large data holdings are at the heart of the potential for some platform markets to be dominated by single players and for that dominance to be entrenched in a way that lessens the potential for competition for the market. In these circumstances, if other solutions would not work, data openness, could be the necessary tool to create the potential for new companies to enter the market and challenge an otherwise entrenched business.' HM Treasury UK Government, 'Unlocking digital competition: Report of the Digital Competition Expert Panel' 13 March 2019, 2.89 Page 75, available: https://www.gov.uk/government/publications/unlocking-digital-competition-report-of-the-digital-competition-expert-panel.

132 BTech, Access to remedy and the technology sector: basic concepts and principles, available: https://www.ohchr.org/Documents/Issues/Business/B-Tech/access-to-remedy-concepts-and-principles.pdf.

133 Access to remedy and the technology sector: 'a remedy ecosystem approach' OHCHR 2020 https://www.ohchr.org/Documents/Issues/Business/B-Tech/access-to-remedy-ecosystem-approach.pdf

100. Social media service providers should consult the UNHCHR B-Tech report 'Access to remedy and the technology sector: understanding the perspectives and needs of affected people and groups'[134].

Enforcement of National Criminal Law

101. Groups representing victims of hate crimes raised concerns that even in developed markets such as the UK, social media companies were not complying swiftly with legitimate requests from law enforcement authorities.

102. Social media providers make decisions about the quantity and quality of resources they employ in services such as in investigation and remediation of potential human rights adverse – or even illegal – impacts. Providers should therefore ensure the adequate frameworks are in place to process law enforcement requests expeditiously, and resource them adequately.

103. This does not mean automatically handing over information without consideration of the legitimacy of the request, or considering the privacy of the users involved, but that the question is given appropriate and speedy attention by appropriately qualified staff, bearing in mind the general principle that applicable laws (which themselves respect international human rights laws) should be respected.

Education and Training

104. Groups representing victims of hate crime told us that a simple lack of staff training at global media companies was a factor in harm to minorities. The Interpretative Guide to the UNGPs stress repeatedly[135] the need for adequate staff skills in corporations to fulfil their duty to respect human rights.

105. Social media service providers that have chosen to offer their service globally sometimes forget to ensure that moderators have been trained in the multitude of local issues of persecution of minorities that might arise in markets far away from the corporate headquarters. They should also ensure that training and resources are kept up-to-date.

134 'Access to remedy and the technology sector: understanding the perspectives and needs of affected people and groups' UNHCHR January 2021 https://www.ohchr.org/Documents/Issues/Business/B-Tech/access-to-remedy-perspectives-needs-affected-people.pdf.

135 See Q31 in response to UNGP 17: "It is important for all enterprises to ensure that the personnel responsible for human rights due diligence have the necessary skills and training opportunities.' Or in response to UNGP 19: "Can we build scenarios or decision trees for action across the company so that we are prepared to respond to the most likely or severe potential impact? Do staff need training and guidance on these issues?" Or in response to dealing with conflicting requirements Q83 "the more an enterprise has embedded respect for human rights into its values and the more it has prepared its personnel for ethical dilemmas, through training, scenarios, lessons learned, decision trees and similar processes, the more likely it will be able to identify appropriate and timely responses" OHCHR Interpretative Guidelines (n 8).

106. The trusted flaggers referred to in Guideline 5 are civil society groups who provide expert victim focussed advice on harm arising from social media to the companies who run networks, often with a special access channel to ensure their voice is heard, and can be seen as a mechanism as envisaged in UNGP18 to allow social media providers to draw on external, sometimes specialist, expertise. They can be invaluable to social media service providers but must not be abused as a source of free labour.

Vigilance over Time

107. The responsibility of companies to respect human rights is a continuous one. Where companies choose to offer services that move fast, they put upon themselves an obligation of equally fast-moving risk assessment. Where society is also changing fast (perhaps in part because of the companies' services) then the obligation is increased. This is a basic cost of doing business.

108. Risk management and mitigation should proceed in lock step with software and societal changes. This does not just include hate speech risk-assessments of new features but continuing to risk assess for hate speech in the use or abuse of speech the use of older features.

109. Social media services evolve almost constantly just to stand still and stay secure - software updates in the biggest services delivered on a continuous flow basis.[136] New features are software shipped frequently. The world of people using the software changes even faster – especially on platforms with global reach.

110. National level governance frameworks for social media are also evolving rapidly.

111. Lesson learned in the implementation of this code should be fed back into a service providers' compliance function and to the Special Rapporteur for Minority Issues.

Towards Guidelines

112. The discussion in this paper has led us to suggest, tentatively a set of draft guidelines for the Special Rapporteur to consider. These are set out in the Annex below.

136Rapid release at massive scale' Facebook 31 August 2017, available: https://engineering.fb.com/2017/08/31/web/rapid-release-at-massive-scale/.

Annex A – Draft guidelines

Guideline 1: Responsibility, Risk Assessment, Mitigation and Remediation

(1) Social media service providers should have a policy commitment to take action to combat hate speech arising on their service. This commitment should be endorsed by the global board and all 'c-suite' executives.

(2) Social media service providers should carry out a suitable and sufficient assessment in relation to each nation in which the social media service is used as to the risk of harm from hate speech attacks on people or groups based on their identity arising from the operation of the service or any elements of it. The risk assessment should be accompanied by a mitigation plan that addresses at least the issues raised later in these Guidelines.

(3) The risk assessment should, in particular, be carried out before the launch of any new service, any new feature, or any service or feature is made available in any new nation.

(4) Service providers should identify metrics to assess the appropriateness and success of the mitigation plan and use them to assess effectiveness of the mitigation plan regularly (at least annually) and revise the mitigation plan accordingly.

(5) The risk assessment should be reviewed by the service provider on an ongoing basis or, if there is reason to suspect that it is no longer valid; or there has been a significant change in the matters to which it relates; and where as a result of any such review changes to a mitigation plan are required the service provider should make them.

(6) Risk assessments and mitigation plans should be recorded, retained for a period of not less than three years and published on the service provider's website in an accessible manner in languages commonly used on the service.

(7) All measures taken in the following guidelines should feed back into the risk assessment as it evolves.

Guideline 2: Safety by Design

(1) Social media service providers should implement appropriate "safety by design" technical and organisational measures including but not limited to those detailed in these Guidelines to minimise the risks of those harms arising from hate speech and mitigate the impact of those that have arisen, taking into account the nature, scope, context and purposes of the online platform services and the risks of harm arising from the use of the service.

Guideline 3: Access to the social network and content creation

(1) Social media service providers should have in place Terms of Service which are visible and understandable by all likely users. This includes providing different language versions of the Terms of Service appropriate to the states in which the service is made available. The Terms of Service must be visible to would-be users before they sign up to the service. The Terms and Conditions must be fit for purpose taken against the provider's values. Community standards should also be visible and should, where relevant, cover the content of advertising including policies concerning advertising sales in respect of promoting harmful content or for malicious intent in respect of members of minority communities.

(2) Social media service providers should ensure and be able to demonstrate that their sign-up processes have taken an appropriate, proportionate approach to the principle of "knowing your client" (KYC), both in relation to users and in relation to advertisers.

(3) Social media service providers should risk assess the tools for the creation of content – this includes but is not limited to bots (including chatbots), bot networks, deep fake or audiovisual manipulation materials and content embedded from other platforms and synthetic features such as gifs, emojis, hashtags.

(4) Social media providers should undertake regular, systemic reviews of their Terms of Service and Community Guidelines to ensure that they remain up to date, effective, and proportionate.

Guideline 4: Discovery and Navigation

(1) Social media service providers should review their recommender systems, especially their automated systems, so that they do not cause foreseeable harm through promoting hateful content, groups or other users to follow for example by rewarding controversy with greater reach, causing harm both by increasing reach and engagement with a content item.

(2) Social media service providers should consider whether the recommendation of "counter speech" is effectively supported by their systems;

(3) Social media service providers should consider the impact of autoplay functions, especially in the context of content curated or recommended by the provider. Where the service provider seeks to take control of content input away from the person through autocomplete or autoplay (see below) the provider should consider how this might affect a person's right to receive or impart ideas.

(4) There should be due consideration of the circumstances in which targeted advertising may be used and oversight over the characteristics by which audiences are segmented.

(5) Social media service providers should consider the need for explainability or interpretability, accountability and auditability in designing AI/ML systems.

Guideline 5: User Response, User Tools

(1) Social media service providers should consider what tools, in addition to content and behaviour reporting tools, are necessary to allow users to improve their control of their online interactions and to improve their safety. These could include:

(a) controls over recommendation tools, so a user can choose for example to reject personalisation;

(b) user-set filters (over words or topics);

(c) tools to limit who can get in touch/follow a user, or to see a user's posts;

(d) tools to allow users to block or mute users, or categories of user (eg anonymous accounts);

(e) Controls for the user over who can and cannot redistribute their content or user name/identity in real time; and

(f) The ease of use of these tools and their prominence such that users are aware they exist

2) Service providers should have reporting processes that are fit for purpose in protecting members of minority groups from hate speech and wider harms, that are clear, visible and easy to use and age-appropriate[137] in design. Thought should be given to reporting avenues for non-users.

3) Service providers should have in place clear, transparent, fair, consistent and effective processes to review and respond to content reported as hate speech. Users must be given the ability to submit third-party content to the companies' intelligence systems in relation to specific cases of content violation.

(4) A platform provider should consider the speed and ease of transmission, for example methods to reduce the velocity of forwarding and therefore cross-platform contamination.

137 UNHCR: General comment No. 25 (2021) on children's rights in relation to the digital environment CRC/C/GC/25 Available here: https://tbinternet.ohchr.org/_layouts/15/treatybodyexternal/Download.aspx?symbolno=CRC%2fC%2fGC%2f25&Lang=en

Guideline 6: Moderation

1) Social media service providers should have in place expanded guidance explaining their policies (and how these are developed, enforced and reviewed, plus the role of victims' groups and civil society in developing them) on hate speech towards members of minority communities in each country in which they operate. Guidance should include what activity and material constitutes hateful content, including that which is a hate crime, or where not necessarily illegal, content that may directly or indirectly cause harm to others. This includes: abuse, harassment and intimidation; hate speech; content promoting hostility or incitement to hatred based on legally protected characteristics whether in isolation or an intersectional manner; and disinformation where this creates the promotion of hostility or incites hatred.

2) Social media service providers should have in place sufficient numbers of moderators, proportionate to the service provider size and growth and to the risk of harm who are able to review harmful and illegal hate speech and who are themselves appropriately supported and safeguarded.

3) Social media service providers should have in place disaggregated notification systems for each type of hate speech towards minority groups in each country in which its service is available and operates to ensure the correct moderators, trained in their specialist subjects and on related language and cultural context considerations (where proportionately reasonable), are able to review the appropriate content, and for transparency purposes. The typology should be developed with victim representatives.

4) Social media service providers should have in place processes to ensure that where machine learning and artificial intelligence tools are used, they operate in a non-discriminatory manner and that they are designed in such a way that their decisions are explainable and auditable. Users should be informed of the use of such tools. Machine learning and artificial intelligence tools cannot wholly replace human review and oversight. `

5) Social media service providers should when receiving a notification of hate speech, review such a report taking into account national laws, their compliance with international human rights standards and the Terms of Service where the comment is made or where it is directed.

6) Social media service providers should have clear timeframes for action against content that is illegal or which is contrary to the provider's terms of service. Awareness begins at the time flagged content, by means of email, in-platform notification or any other method of communication, is received.

7) Social media service providers should take action, proportionate to risk, on content which is not deemed to be illegal but is considered to break their Terms of Service or Community Guidelines as soon as it is identified. Acceptable actions on a piece of content which violates a provider's Terms of Service can include –

(a) Label as inaccurate/misleading/contrary to the rules;

(b) Demonetise content;

(c) Suppress content in recommender tools;

(d) (Removal of content;

(e) Termination of account;

(f) Suspension of account;

(g) Geo-blocking of content;

(h) Geo-blocking of account;

(i) A strike, if a strike system is in place.

8) Social media service providers should have systems of assessment and feedback to the initial reporter and the owner of content that has been flagged and actioned to ensure transparency of decision making. Users should be kept up to date with the progress of their reports and receive clear explanations of decisions taken.

9) Social media service providers should put in place a right of appeal on all decisions made concerning illegal or harmful content, or content that has been flagged as illegal or harmful content. All users must be given a right to appeal any measures taken against them (see para 7), whether in full or in part. Users must be able to present information to advocate their position.

10) Social media service providers should acknowledge an appeal request, within 24 hours of receipt. If more time is needed to assess the content, the user must be informed.

11) Social media service providers should have appeals systems which must take no longer than seven days to assess appeals, except in exceptional circumstances. Exceptional circumstances could include a major disaster, or an event or incident of the same magnitude.

12) Social media service providers should consider the need to have intelligence systems for investigating harms organised off-platform for attack of users who are members of a minority community on a given platform and whether to share such intelligence, when received, with other platforms.

13) Social media service providers should consider putting in place an appropriate trusted flagger programme that maintains independence from the service provider and from governments. The programme must include nongovernment organisations and other experts, who will be vetted, to inform on policy development and report on new

trends in harmful and illegal content. In order to ensure an effective working relationship with members of Trusted Flagger programmes, service providers should:

(a) Ensure trusted flaggers are not used as a sole provider of flagging content;

(b) Ensure trusted flaggers are appropriately compensated and incentivised for work provided to companies to ensure their compliance while not compromising their independence and impartiality;
(b) Hold regular meetings (with members of the trusted flagger programmes) to review content decisions and discuss any concerns;

(c) Provide support for trusted flaggers who are exposed to harmful content, as per the support provided to the companies' own moderators, whether directly employed or working for out-sourced companies.

Guideline 7: Safety Testing

1) As part of their risk assessment and mitigation processes, social media service providers should carry out or arrange for the carrying out of such testing and examination of their systems as may be necessary to carry out due diligence in reducing harms arising from attacks on minorities, bearing in mind respect for the human dignity of people involved or affected by those tests, as well as ethical considerations relating to experiments involving human participants.

2) Testing should specifically include recommendation and curation functions and automated curation and moderation systems.

Guideline 8: Supply Chain Issues

(1) Social media service providers which outsource any part of their business, including moderation of content, applications, GIFs, images, or any other content or tools, including safety tech, should ensure the vendor adheres to the social media provider's Terms of Service and Community Standards and that they have employee and mental health protection policies in place that adhere to the same standard.

2) Processes should be in place for users to report content or tools provided by a vendor which is illegal or violates the service provider's Terms of Service or Community Standards and Guidelines.

(3) Social media service providers should ensure adequate information is available to the vendors on their Terms of Service and Community Guidelines to pre-empt any violations.

Guideline 9: Victim Support and Remediation

(1) Social media service providers must take steps to ensure that users who have been exposed to hateful material are directed to, and are able to access, adequate support. Support can include –

(a) Signposting and access to websites or helplines dealing with the type of hatred viewed by the user or witnessed by others who may be affected by the content, even if not the designated target;

(b) Information from, and contact details for, services providing victim support or mental health support after being exposed to hateful and harmful materials;

(c) Strategies to deal with being exposed to hateful material.

Guideline 10: Enforcement of National Criminal Law

(1) Social media providers must have in place a point of contact for law enforcement authorities for each nation in which the service operates. The contact is responsible for giving information about illegal content to law enforcement authorities under para 2. This includes –

(a) Information about the content;

(b) The details of the user, including location;

(c) Details of enforcement action on the content undertaken by the provider; and

(d) Other materials relevant to criminal investigations.

(2) Information requested by government and law enforcement authorities in accordance with local law should be delivered within the time frame specified by national rules or no later than one month of receiving the request. In exceptional circumstances this can be extended, with written approval from the relevant authorities placing the request, with a full expected time frame set out.

(3) effective protections should be put in place by social media providers to ensure flagging and court orders are not used for malign purposes by Government agencies or law enforcement of any kind to remove content they find objectionable, which is neither illegal nor harmful.

Guideline 11: Education and Training

(1) Social media service providers should put must consider putting in place appropriate, updated education and training on hate speech for all staff and subcontractors involved in the content production and distribution chain. This includes senior executives, designers, developers, engineers, customer support and moderators, designed in consultation with independent Trusted Flaggers to insure diversity and inclusion in respect of each states.

(2) Materials used for training on illegal and harmful content must be made available to the Government, any Regulator, law enforcement authorities and Government agencies upon lawful request.

(3) Within the service itself providers should ensure that training and awareness tools are readily available to users on the Terms of Service and Community Guidelines to ensure users are aware of permitted content and behaviours on the platforms.

Guideline 12: Vigilance over Time

(1) Social media service providers must have plans for ongoing review of their efforts in tackling hate speech. This might include engagement with relevant experts or organisations to advance policy development. The providers shall adapt internal processes accordingly, to drive continuous improvement and in particular shall regularly review and update when appropriate technical and organisational measures implemented under this code.

Carnegie UK
Andrew Carnegie House
Pittencrieff Street
Dunfermline
Fife, Scotland
KY12 8AW

T +44 (0)1383 721445
www.carnegieuk.org

Registered Charity No: SC 012799 operating in the UK
Registered Charity No: 20142957 operating in Ireland

collective wellbeing
Carnegie UK