CarnegieUK TRUST

CHANGING MINDS • CHANGING LIVES

DRAFT Code of Practice in respect of Hate Crime and wider legal harms: covering paper

June 2021

Introduction and background

1. There are numerous codes of practice in place for social media companies. They are almost all voluntary in nature, limited in scope and have little to no impact in practice. Often, companies argue that their own activities go far beyond stated codes of practice which quickly b*ec*ome unfit for purpose. The initial consultation offerings from Government for the (then) Online Harms Bill indicated that a statutory Duty of Care would be introduced for social media platforms, or those allowing the sharing and distribution of user-generated content, and that Codes of Practice which informed the Duty of Care would be published.[1] A newly proposed regulator would, as one part of its wider examinations, assess adherence to the codes and judge failures to apply the Duty of Care accordingly. Codes on Terrorism and Child Sexual Exploitation and Abuse have been given prominence and special status, with oversight from the Home Secretary; drafts of these have since been published. The 'Hate Crime' Code of Practice has not been given such status and does not, as proposed, include wider harms.

2. That is why we have worked over the course of a number of months with a number of civil society organisations, who speak to the lived experience of many groups that experience hate crime online, on the development of this model Code of Practice for Hate Crime and wider legal harms.[2]

[1] The Online Harms White Paper (2019) provided examples of the codes of practice a regulator might draw up for companies to fulfil their duty of care in the following areas: CSEA and terrorist use of the internet (these voluntary codes, drafted by the government, were published with the Full Response to the White Paper consultation in December 2020); Serious Violence; Hate Crime; Harassment; Disinformation; Encouragement of self-harm and guidance; Online Abuse of public figures; Interference with legal proceedings; Cyberbullying; Children Accessing Inappropriate Content. (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf pp64-76)

[2] This draft Code has been developed with the invaluable 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 has also benefited from the insight and feedback on the draft Code from further contributors, including representatives of the major tech platforms, central government policymakers and regulators, at a workshop hosted by Carnegie UK Trust on 26th February 2021.

A draft of this Code was discussed at a workshop in February 2021, which brought together those initial collaborators with representatives of the major platforms, policymakers and regulators. The headlines from their feedback are set out below and provide important framing for consideration of the Code itself as well as an indication of the different perspectives that we expect to encounter in this next stage of engagement and advocacy.

3. The Code draws on, and should be read in conjunction with, the Codes and other references in the annex, including the European Code of Conduct on Illegal Hate Speech Online and the Digital Economy Act Code of Practice. It offers systems-level solutions to addressing harms and could form the basis of a regulatory Code in this area. It is aimed at outlining principles rather than an exhaustive set of rules and service operators should aim to engage with the spirit of those principles and not look just to the letter of the Code. This draft Code does not at this stage synchronise fully with the Government's draft Online Safety Bill[3]. Furthermore, the work needs to be carefully considered in respect of devolved administrations and powers.

4. The challenges of discussing the merits and content of a draft Code without seeing the "parent" legislation to which it will be attached was a recurring theme at our February workshop. However, we did not wish to wait for that draft Bill in order to revise the Code "to fit" the Government's framework. With pre-legislative scrutiny about to start, we hope now to use this draft to start the debate on the need for the Government to include such a Code in the first place and also to re-emphasise the importance of a systemic (rather than content regulation) approach, from which a Code such as this could then naturally flow. While the draft Bill draws a boundary between the treatment of content that is likely contrary to the criminal law and that which is not, we still think there are merits in this holistic approach for two reasons. First, it demonstrates that not all instances of hate speech necessitate the same response and that in this we are not limited to take down. Secondly, it deals with the point that such speech can escalate and, crucially, minor-sounding instances can be linked to a larger picture of harm.

Commentary on the draft Code

5. Our workshop participants agreed that the strengths of the Code lay both in the content and in the means of its development. The systemic approach, with a focus on design and processes

[3] https://www.gov.uk/government/publications/draft-online-safety-bill

drawing on Carnegie's model[4], is positive. The Code builds on good practice but raises the bar, and the level of detail shows how such an approach would work effectively in practice. Keeping the emphasis on principles and outcomes will be important, as well as keeping the wording broad to take account of things that haven't yet emerged or not been thought of yet.

6. There is a "cart before horse" challenge (noted above) that arises because – at the time of our workshop – the Government's draft Bill had not yet been published at the time of our discussions. The development of the Code – ahead of the publication of the Government's legislative proposals – was seen as both a strength and risk. While it allows a truly systemic, process-led approach to be set out for debate and engagement, it was not clear how successfully it would integrate with the overarching legislation – and now, in the light of the draft Bill, that remains a challenge.

7. However, feedback suggested that the strength of this Code lay in the way it set out what could be seen as an overarching systems Code which could apply to other harms. (For example, one would not necessarily need the repetition of many of the clauses between multiple codes if there was an overarching systems code from which they flowed)**. One area for the Government and Parliamentarians to consider as it responds to scrutiny of the draft Bill is whether there is an opportunity to create a single core set of practices that would apply to all companies and then set addendums for specific harms and for specific companies**.

8. The balance between illegal activity and legal but harmful was judged to be about right. Workshop participants judged that it was important that the latter was included, particularly as there is a risk that the legislation may not give enough prominence to it; we are still working to understand the draft Bill on this aspect. The draft Code was also deemed to align well with international developments and best practice.

9. There were differences of opinion – inevitably – between civil society and tech company views on the necessary level of prescription and specificity in the Code. The former (in general) viewed the detail as a necessary baseline to raise the bar and ensure all companies comply, without introducing too much wriggle room for dilution of the requirements; the latter (representing the

[4] See, for example, our April 2019 reference paper (https://d1ssu070pg2v9i.cloudfront.net/pex/carnegie_uk_trust/2019/04/08091652/Online-harm-reduction-a-statutory-duty-of-care-and-regulator.pdf) and our draft Online Harm Reduction Bill (https://www.carnegieuktrust.org.uk/publications/draft-online-harm-bill/)

bigger platforms) would prefer less detail and greater flexibility to enable companies to develop solutions in a way that suits their specific circumstances.

10. While a concern was raised that the Code might treat all companies as the same, the emphasis on proportionality was welcomed and could be brought out further, particularly how it will work in practice. It was judged that a focus on proportionality will ensure that burdens on smaller companies will be minimised and, in the next stage of iteration, it will be important to consult more widely with them as well as the big platforms. That said, in terms of the detail and prescription, there was a view that the Code should not become too static: it needs to remain dynamic and flexible, to adapt to a rapidly changing landscape, and to remain focused on outcomes (which would remove some of the issues that companies might face if their particular systems and processes were not relevant for parts of the code).

11. Regardless of the systemic approach of this Code (and the opportunity it provides to set out a "template" for an overarching systems Code from which others might flow), the risk of duplication and a proliferation of codes is real issue: while the approach of the hate crime Code is complementary to those already published by the Government (on terrorist content and CSEA), it is not clear how they might fit together (or the respective status of codes, guidance, principles, etc) and plans to introduce further codes for individual harms may prove confusing and burdensome for companies, with competing requirements and different responses. There are also some boundary issues between this Code and the Code on terrorism.

<u>Specific areas of focus</u>

12. The **enforcement approach** was agreed in principle by participants at our workshop, but this needed to be balanced against safeguards. In particular, the proposals about data capture and passing data to law enforcement raised concerns, particularly around the requirement that platforms should make decisions about what content should be passed to criminal investigations and/or the government. The Code should align more closely with the existing processes that companies have for engagement with law enforcement agencies. The good Samaritan clause was also well balanced with due diligence.

13. The **information-gathering powers** are important – it is vital for civil society and regulators to understand better what is going on (eg banned accounts being set up again). Access to data is key

to transparency between regulators and platforms. A focus on **metrics** was important (and would be headline grabbing) but it is important to get into the incentive models and the nuances around this. Too much specificity re metrics will mean that companies might game the system by focusing on numbers rather than why that particular thing is important, and could incentivise the worst behaviours.

14. The clauses on **trusted flaggers** were important: these roles bring more understanding of the communities affected by hate crime than the platforms but there needed to be more clarity on what these programmes should look like, labour and resource demands and the clarification of accountability and liability.

15. The collaborative, iterative approach facilitated by Carnegie UK Trust was welcome and was seen as a strength. Participants felt that Ofcom should be encouraged to set out how it will meaningfully engage with civil society and how it will improve on companies' current processes for this to ensure that all relevant players are in the room for similar discussions.

Next steps

16. As noted above, this draft Code does not at this stage synchronise with draft Online Safety Bill, which itself will be subject to modification and revision as it passes through the Parliamentary scrutiny process. Instead, the Code contains many clauses that set out our own approach to a truly systemic, risk-managed regulatory regime, built on an overarching statutory duty of care (rather than multiple duties, as is the Government's intent). As such, it is a commentary on the more limited approach the Government has adopted, which we discuss further in our full response. We are publishing it now to start the debate and to improve its design.

17. **We would welcome feedback on the draft by 30th June 2021 and will take this into account as we work to revise the draft to dock fully with the draft government Bill.**

Please submit feedback to by 30th June to: info@carnegieuk.org

Draft Code of Practice in respect of Hate Crime and wider legal harms

Risk Management and Assessment

(1) Companies shall have carried out a suitable and sufficient assessment of the risk of harm arising from attacks on those with protected characteristics, people under 18 and vulnerable people arising from the operation of the service or any elements of it. This includes annual risk assessment reviews as well as emergency processes to address new or emerging risks. The issues to be covered in the assessment should be set down by the Regulator; any guidance as to format and evidence should be taken into account. The assessment shall be reviewed by the operator 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 an assessment are required the operator shall make them. Such assessments shall be recorded and retained for a period of not less than three years or as set out by the regulator in guidance.

(2) Companies shall implement appropriate "safety by design" technical and organisational measures including but not limited to those detailed below to minimise the risks of those harms arising 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;

(3) Companies shall carry out or arrange for the carrying out of such testing and examination as may be necessary for the performance of the duty of care in respect of harms arising from attacks on those with protected characteristics, 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;

(4) Companies shall regularly review and update when appropriate technical and organisational measures implemented under this code. Companies shall also keep the

appropriateness and effectiveness of such measures under review during the period the online platform service is offered.

(5) Companies should ensure and be able to demonstrate their systems are safe by design, including addressing the following concerns;

(a) That algorithms do not cause foreseeable harm through promoting hateful content, for example by rewarding controversy with greater reach, causing harm both by increasing reach and engagement with a content item;

(b) That speed of transmission has been considered, for example methods to reduce the velocity of forwarding and therefore cross-platform contamination;

(c) Use of tools by actors creating or spreading harms against those with protected characteristics in order to cause harm, which are able to operate owing to weak platform polices on enforcement and moderation. This includes but is not limited to bots, bot networks, deep fake or audio-visual manipulation materials and content embedded from other platforms;

(d) An appropriate approach to the principle of knowing your client [KYC] to address harms spread by those using false or anonymous identities;

(e) Consideration of the circumstances in which targeted advertising may be used and oversight over the characteristics by which audiences are segmented;

(f) Systems for cross-platform co-operation to ensure knowledge about repeat offenders that may present a foreseeable risk of harm in relation to attacks of those with protected characteristics;

(g) Use of tools including but not limited to prompts which clarify an individual's intended search;

(h) Policies concerning advertising sales in respect of promoting harmful content or for malicious intent in respect of those with protected characteristics.

Enforcement

(1) Companies should have methods to proactively identify content or activity constituting criminal hate speech, to either prevent it being made publicly available or prevent

further sharing.

(2) Companies must have in place Terms and Conditions which are clear, visible and understandable by all likely users. The Terms and Conditions must also be fit for purpose for their compliance with the statutory duty of care.

(3) Companies must have reporting processes that are fit for purpose in respect of hate crime and wider harms, that are clear, visible and easy to use and age-appropriate in design. Thought should be given to reporting avenues for non-users.

(4) Companies must have in place clear, transparent and effective processes to review and respond to content reported as illegal and harmful.

(5) Companies must have in place effective and appropriate safeguards in full respect of fundamental rights, freedom of expression and relevant data protection regulation. This includes, specifically, taking reasonable steps to ensure users will not receive recommendations to criminal, hateful or inappropriate content.

(6) Companies must have in place Community Guidelines explaining their policies (and how these are developed, enforced and reviewed, plus the role of victims' groups and civil society in developing them) on harmful content, including 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:

(a) Terrorist content;
(b) Child Sexual Exploitation and Abuse (CSEA) content;
(c) Abuse, harassment and intimidation;
(d) Stalking;
(e) Hate speech;
(f) Content promoting hostility or incitement to hatred based on legally protected characteristics whether in isolation or an intersectional manner;
(g) Disinformation where this creates the promotion of hostility or incites hatred based on legally protected characteristics;

(h) Criminal activity.

(7) Companies must have in place sufficient numbers of moderators, proportionate to the company size and growth and to the risk of harm who are able to review harmful and illegal content and who are themselves appropriately supported and safeguarded. Machine learning and Artificial Intelligence tools cannot wholly replace human review and oversight. Companies 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.

(8) Companies must have a disaggregated notification system for each type of harmful and illegal content 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.

(9) When receiving a notification of content, a Company must review such a report taking into account national laws and the Terms of Service.

(10) A company must remove content that has been deemed to be illegal within 24 hours of becoming aware of such content. Awareness begins at the time flagged content, by means of email, in-platform notification or any other method of communication, is received.

(11) A Company must 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 and no later than 24 hours]. Acceptable actions on a piece of content which violates a Company's Terms of Service can include –

(a) Removal of content;
(b) Termination of account;
(c) Suspension of account;
(d) Geo-blocking of content;

(e) Geo-blocking of account;
(f) A strike, if a strike system is in place.

(12) Without prejudice to (10), companies must have clear guidelines as to what constitutes an expedient time frame for the removal of (or temporarily limiting access to) hateful content and comply with them.

(13) Companies must have systems to prevent and identify those abusing and misusing services to create harm, including persistent abusers across a range of harms and those using anonymous accounts to abuse others.

(14) Companies must put in place 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.

Outsourced Content

(1) Companies that outsource any part of their business, including moderation of content, applications, GIFs, images, or any other content or tools, must ensure the Vendor adheres to the Terms of Service and Community Guidelines of the company and that they have employee and mental health protection policies in place that adhere to the same standard.

(2) Processes must be in place for users to report content provided by a Vendor which is illegal or violates the Company's Terms of Service or Community Guidelines.

(3) Companies must ensure adequate information is available to the Vendors on their Terms of Service and Community Guidelines to pre-empt any violations.

Right of Appeal

(1) Companies must 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.

(2) All users must be given a right to appeal any termination of service, suspension, geo-blocking or removal of content, whether in full or in part. Users must be able to present information to advocate their position.

(3) Companies are to ensure that Protected Speech is not removed from the platform unduly.

(4) Companies must acknowledge an appeal request, within 24 hours of receipt. If more time is needed to assess the content, the user must be informed.

(5) Appeals must take no longer than seven days to assess, except in exceptional circumstances. Exceptional circumstances could include a major disaster, or an event or incident of the same magnitude.

User Support

(1) Companies should provide advice and tools for "digital self-care" such that users can take steps to protect themselves in the first instance from exposure to hateful content and that these are built into new features.

(2) Companies 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.

Transparency

(1) Companies must engage in regular self-auditing to ensure compliance with the Code of Conduct. Companies must provide quarterly transparency reports, based on these audits, of content removed which can viewed in the United Kingdom, regardless of the content's origin or origin determined by IP address. Thought should be given to effective ways to communicate with users about these reports.

(2) Transparency reports must include data and statistics produced by the IT companies. The reports must include information on –

(a) Content removal;

(b) Content removal, disaggregated by removal reason;

(c) Content removal by first source of detection. This must include both automated flagged and human flagging;

(d) Removal times, disaggregated by removal times;

(e) Appeals, disaggregated by reason for reinstatement;

(f) Content that has been reinstated;

(g) Law enforcement requests for removal;

(h) Court orders for content removal and adherence to court orders;

(i) Number of requests for removal by the Regulator;

(j) Regulator interventions.

(3) Companies must have in place the ability to grant independent scholars, academics, researchers and others with genuine, verifiable research interests that are independent of the company, the ability to access anonymous data in order to better understand the situation of illegal and harmful content, adherence to the Code of Conduct and other online activity.

(4) Companies must be able to provide information about the prevention and identification of abuse and misuse of services, including persistent abusers across a range of harms; and those using anonymous social media accounts to abuse others. This must be provided in

full to government and law enforcement agencies upon request within one month of receiving the request.

Governance and Authority

(1) IT Companies must have in place a point of contact for law enforcement authorities. The contact is responsible for giving information about illegal content to law enforcement authorities. 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 IT Company;

(d) Other materials relevant to criminal investigations.

(2) Information requested by government and law enforcement authorities must be delivered within 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) Protections must be put in place by IT Companies to ensure flagging and court orders are not used for nefarious purposes by Government agencies or law enforcement of any kind to remove content they find objectionable, which is neither illegal nor harmful.

Education and Training

(1) IT Companies must put in place appropriate, updated education and training for moderators, designed in consultation with independent Trusted Flaggers to insure diversity and inclusion.

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

(3) IT Companies must have in place an appropriate, independent Trusted Flagger programme. The programme must include Non-Government 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 10 working relationship with members of Trusted Flagger programmes, IT Companies must –

(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 IT Companies to ensure their compliance while not compromising their independence and impartiality;

(c) Hold regular meetings (with members of the Trusted Flagger programmes) to review content decisions and discuss any concerns;

(d) 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.

(4) IT Companies must provide educational tools and guidelines on their Terms of Service and Community Guidelines to ensure users are aware of permitted content on the platforms.

Planning and Review

(1) Companies must have plans for continually reviewing their efforts in tackling hateful material and adapt internal processes accordingly, to drive continuous improvement. This might include engagement with relevant experts or organisations to advance policy development.

(2) Companies should have intelligence systems for investigating harms organised off-platform for attack of users on a given platform and should actively and regularly share such intelligence, when received, with other platforms.

(3) Users must be given the ability to submit third-party content to the Companies' intelligence systems in relation to specific cases of content violation.

Enforcement of this code of practice

(1) The Regulator will assess whether actions taken by a company to comply with this code of practice are suitable and sufficient to address the risk of harm arising from hate crime in the operation of that company's service. The regulators assessment will be informed by a dialogue with a range of actors including the company concerned, victims of harm, civil society actors, the regulator's own research and any other information that the regulator considers to be relevant.

(2) Should the regulator find that a company has not taken suitable and sufficient steps under this code and harm from hate crime has occurred the regulator will employ its enforcement management model to consider whether sanctions under the online harms regime are required.

Glossary

Flagged: When a piece of content is reported to an IT company.

Harmful Content: Any content which is harmful to users of IT companies and wider British society. This includes activities that incite or engage in violence, intimidation, harassment, threats, defamation or other hostile act based on protected characteristics.

Illegal Content: Any content deemed to be illegal under British Law. This can include harmful content.

IT Companies / Company: Any platform accessible via the internet which allows for user-generated content to be hosted by the platform. Content can include text, imagery, photographs, videos, comments, sound and performances. Companies can include social media platforms, video sharing platforms, event scheduling/ticketing platforms, public chat services or group communications, websites, blogs or message boards.

Moderator: Any person who reviews content for the IT Companies. This can include third party moderators, vendors, to who the IT Companies outsource moderation and review tasks.

Regulator: The Government's online harms regulator.

Trusted flagger: an individual or organisation with particular expertise and responsibilities for the purposes of tackling illegal content online.

Vendor: Any external company, application, tool or other contracted out service which is available to users on the platform. This can include, but is not limited to, images, content moderation.

Further recommended reading:

The European Code of Conduct on Illegal Hate Speech Online

https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

Change the terms: Reducing Hate Online, Recommended Policies

https://www.changetheterms.org/terms

Digital Economy Act Code of Practice https://www.gov.uk/government/publications/code-of-practice-for-providers-of-online-social-media-platforms

Age Appropriate Design Code: https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/age-appropriate-design-a-code-of-practice-for-online-services/

AVMSD/ VSP provisions: https://www.ofcom.org.uk/tv-radio-and-on-demand/information-for-industry/vsp-regulation

The Alan Turing Institute: How much online abuse is there? A systematic review of evidence for the UK: https://www.turing.ac.uk/sites/default/files/2019-11/online_abuse_prevalence_full_24.11.2019_-_formatted_0.pdf