EMC Submission No. 82
Received 30 September 2020

Submission to Electoral Matters Committee; Parliament of Victoria (59th Parliament)

Inquiry into the Impacts of Social Media on Elections and Electoral Administration

This submission presents to the Committee a summary of our work for Carnegie UK Trust in the UK on a statutory duty of care for harm reduction and how it might affect the impacts of social media in general, and the consequences for elections. This submission does not deal with the specifics of rules relating to political advertising in the context of an election. However well those rules work, they will not be sufficient where social media are used to spread misinformation and disinformation.

Background to the Project

Carnegie UK Trust was set up in 1913 by Scottish-American philanthropist Andrew Carnegie to improve the well-being of the people of the United Kingdom and Ireland, a mission it continues to this day. Carnegie particularly charged the trustees to stay up to date and the trust has worked on digital policy issues for some years.

In 2016 Woods and Perrin carried out work with an MP (the private members bill 'Malicious Communications (Social Media) Bill) to try to ensure that social media platforms gave adequate tools to users to help them defend themselves from online abuse. This focus on design features and tools formed the basis for a larger project that Woods and Perrin commenced in early 2018 after the UK Government's Internet Safety Strategy Green Paper in Autumn 2017 detailed extensive harms but few solutions. Initially published as a series of blogs, the work developed into a public policy proposal to improve the safety of users of internet services through a statutory duty of care, enforced by a regulator[1]. A full reference paper[2] drawing together their work on a statutory duty of care was published in April 2019, just prior to the publication of the UK Online Harms White Paper[3].

This work has influenced the recommendations of a number of bodies in the UK including select committees in the UK Parliament, charities and the UK Chief Medical Officers.[4] More broadly, Woods

1 https://www.carnegieuktrust.org.uk/project/harm-reduction-in-social-media/

2 https://d1ssu070pg2v9i.cloudfront net/pex/carnegie_uk_trust/2019/04/08091652/Online-harm-reduction-a-statutory-duty-of-care-and-regulator.pdf

3 https://www.gov.uk/government/consultations/online-harms-white-paper

4 https://www nspcc.org.uk/globalassets/ documents/news/taming-the-wild-west-web-regulate-social-networks.pdf; https://www.childrenscommissioner.gov.uk/2019/02/06/childrens-commissioner-publishes-astatutory-duty-of-care-for-online-service-providers/; https://www.gov. uk/government/publications/uk-cmo-commentary-on-screen-time-and-social-media-map-ofreviews/; https://publications.parliament.uk/pa/cm201719/cmselect/cmsctech/822/82202 htm; https://www.parliament.uk/business/committees/committees-a-z/commons-select/digitalculture-media-and-sport-committee/news/immersive-technology-report-17-19/; https://labour.org.uk/press/tom-watson-speech-fixing-distorted-digital-market/; https://www.parliament.uk/business/committees/committees-a-z/lords-select/

gave evidence to the International Grand Committee on Fake News; while it did not make specific reference to this work, a report to the French Ministry of Digital Affairs referenced a "duty of care" as the proposed basis for social media regulation.[5] This suggests that the model is not specific to the UK but could be adapted by other jurisdictions.

Systems-based Regulation

Carnegie UK proposed a shift from the regulation of specific items of content to a focus on the design on platforms (including business models and resourcing of complaints systems). This is based on the assumption that design choices can have an impact on the content posted and the way information flows across communications platforms – including but not limited to recommender algorithms. Rather than specify individual rules, which might quickly become outdated both as regards to technologies and services available and the problems faced, we proposed an overarching duty on operators to ensure, so far as possible, that their services were 'safe by design'. Borrowing from the tort of negligence the concept of a duty of care (which as a private law tool has an analogue in many countries), the Carnegie proposal suggested a statutory duty of care that would set down this general obligation to take *reasonable* steps to address *foreseeable* harm. Note, it is not expected that the duty of care will lead to a perfect environment – it cannot solve all problems on the Internet. It may improve the general environment so as to allow more targeted, content focused measures if needed. In this context, we should note also the importance of data protection rules.

In general, the systems-based approach is neutral as to the topics of content. Moreover, most interventions allow speech to continue, but affect is visibility (e.g changes to recommender algorithm/ autoplay switched off), velocity of spread (number of people to whom one message may be forwarded) and – perhaps – manner of expression (reminders as to rules relating to harassment and hate speech). Such interventions are less intrusive as regards freedom of speech.

The obligation has, in essence, four aspects:

- the overarching obligation to exercise care in relation to user harm;
- risk assessment process
- establishment of mitigating measures; and
- ongoing assessment of the effectiveness of the measures.

The regime envisages an independent regulator with a double-role:

- informing and facilitating good practice (e.g. through the drafting of codes or guidance); and
- verification of compliance of the operators with the duty and, where necessary, enforcement.

communications-committee/inquiries/parliament-2017/the-internet-to-regulate-or-not-toregulate/; https://www rsph.org.uk/our-work/policy/wellbeing/new-filters.html

5 http://www.iicom.org/images/iic/themes/news/Reports/French-social-media-framework---May-2019.pdf

Enforcement action should be context specific and proportionate, especially given the fundamental rights in play (including but not limited to freedom of expression).

While the proposal envisaged that the underpinning statute should set out the types of harm, this does not take away from the fact that this is a general duty. The generality is important for two reasons. First, it allows the regime to develop as technology does, as services and the market change and as understanding of risk and harm increases. It is an element of future-proofing the regime – harm is consistent while the technical state of the art advances. Secondly, the general duty allows operators to take into account their respective services and the risk that those services pose to the sorts of user the services have. It also allows the platform operators to bring their technical and service knowledge into the regime. Finally, the fact that there is a general obligation does not mean that statute cannot specify specific obligations within the general duty – for example, the need to have an effective complaints mechanism, obligations of transparency for particular issues, the need to take particular steps with regard to specific types of content (e.g. child sexual abuse and exploitation material). The general obligation acts as a form of basket for any such specific obligation, given coherence and structure to the regime.

Applicability to Elections – The Role of Design Features and Business Choices

We work on the assumption that elections could be affected by the spread of misinformation and disinformation (presumably resulting in voter confusion); the targeting of votes to dissuade them from voting (or encouraging them to vote)[6]; and the abuse and intimidation of public figures. In the latter instance, content may not appear to be directly related to the election. Note that we have not undertaken empirical work to support these assumptions; they are based on concerns and research carried out by others. We propose the need for regulation because there is emergent evidence of the issues sufficient to shift the burden of proving danger on the part of the State, to the need to prove safety on the part of the operators (the precautionary principle[7]). Note also that perfection is not required; reasonable processes to safeguard the system from abuse should be. Yet it seems that 'safety checking' key features of many social media platforms has not been carried out.

The prevalence and impact of disinformation is mediated by the tech platforms themselves – and this is why regulation at the level of platform is appropriate and justifiable. Choices aimed at maximising user engagement and/or revenue may have unfortunate side effects in terms of content prioritised.[8] The operation of recommender algorithms have, it has been claimed, a tendency to promote extreme content

---

6 See e.g. investigation by Channel 4 news: https://www.channel4.com/news/revealed-trump-campaign-strategy-to-deter-millions-of-black-americans-from-voting-in-2016

7 United Kingdom Interdepartmental Liaison Group on Risk Assessment (UK-ILGRA), The Precautionary Principle: Policy and Application, available: http://www.hse.gov.uk/aboutus/meetings/committees/ilgra/pppa.htm

8 See e.g. J Lanier *Ten Arguments for Deleting your Social Media Accounts Right Now* (Henry Holt and Co 2018); S Zuboff Big Other: Surveillance Capitalism and the Prospects of an Information Civilization (2015) 30 *Journal of Information Technology* 75-89, doi:10.1057/jit.2015.5

or content stimulating emotions such as hate/anger.[9] To hold the attention of these groups (so they can be shown more ads and share more content), platform company algorithms help to generate a climate of outrage and sensationalism, normalising what were once extreme views. Moreover, in some business models, disinformation and misinformation which is engaging is financially rewarded. This threatens to distort electoral outcomes, remove transparency from political debate and undermine the public's faith in rational and accountable political decision-making.

Personal data allows the personalisation of content. It has been suggested that the progressive subdivision of the public into ever more precisely defined target audiences traps people in "filter bubbles" to whom the platforms' algorithms then feed a steady diet of similar, or progressively more polarising or extreme, content that reaffirm and entrench pre-existing beliefs.[10] The use of audience segmentation is also used to deliver adverts. The data protection regulator in Spain banned microtargeting during elections[11], but the decision was quickly quashed. Microtargeting – because it means that users do not see the same information and do not know that others are seeing different adverts– may interfere with public debate around topics of importance. This arguably interferes with the freedom 'to seek, receive and impart information' in Article 19 ICCPR. There are questions around how the platforms determine which types of categorisation is permissible, as well as whether those categories are acceptable (is it appropriate that a voter should be targeted by race, or that psychographic segmentation should be used?). Does the platform operate any form of KYC ("Know Your Client") processes on advertising clients as to who they are (including links to other organisations) and what they advertise? This is particularly important when adverts are used not to inform but to manipulate whether the electorate votes. The UK DCMS select committee noted:

> *Democracy is at risk from the malicious and relentless targeting of citizens with disinformation and personalised 'dark adverts' from unidentifiable sources, delivered through the major social media platforms we use every day. Much of this is directed from agencies working in foreign countries, including Russia. The big tech companies are failing in the duty of care they owe to their users to act against harmful content, and to respect their data privacy rights. Companies like Facebook exercise massive market power which enables them to make money by bullying the*

---

9 See e.g. Zeynep Tufecki, 'You Tube, the Great Radicalizer', New York Times, 10 March 2018, available: https://www nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical html?smid=pl-share; Avaaz, Anti-recism Protests: Divisive disinformation narratives go viral on Facebook, racking up over 26 million estimated views, 12 June 2020, available: https://secure.avaaz.org/campaign/en/anti_protest_disinformation/

10 While some studies (e.g. Bakshy et al 'Exposure to ideologically diverse news and opinion on Facebook' (2015) 348 *Science* 1130, DOI 1-.1126/science.aaa1160) suggest that user choice may be part of this, others have suggested that algorithmic amplifcation has a role to play through the creation of a variant of feedback loop: A. J. B. Chaney et al 'How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility' (2018) RecSys '18, October 2–7, available: https://arxiv.org/pdf/1710.11214.pdf

11 See: Spain: DPA limits the use of data in political campaigning 25 March 2019 https://www.gdprtoday.org/spain-dpa-limits-the-use-of-data-in-political-campaigning/

smaller technology companies and developers who rely on this platform to reach their customers. These are issues that the major tech companies are well aware of, yet continually fail to address.[12]

Some disinformation and misinformation is spread by humans, sometimes deliberately, sometimes inadvertently. In terms of design features, those that ensure the ease of 'sharing' or forwarding material may add to the problem in that people may post without necessarily having read the content or thought whether it is plausible. Some have suggested that counter information strategies and correction of demonstrable false statements should be part of a strategy to counter fake news[13]; insofar as correction works, it needs to reach the same people who saw the disinformation, speedily and with the same emphasis. Fact check labels need to be unmistakeable and easily visible (and security around mechanisms such as trusted users must be tight). The role of bots and coordinated networks of accounts should also be recognised[14]; it is questionable whether all networks recognise risks from bot networks, or take steps to stop bots from using their networks.

In all it can be seen that there are numerous points in the system design and business model where 'architectural features' affect content that is seen by large numbers of users, and that in some instances these features have been deliberately weaponised by certain groups of users. These are not issues relating to content, but to the system itself for which a platform should take responsibility and, at the least, take steps to ensure that they cannot be (easily) weaponised and are otherwise reasonably safe. In this, the onus should be on the platform to document the concerns it has identified and addressed, and demonstrate its compliance with the duty.

Further information

We would be happy to provide further information or discuss this work with the Committee, if helpful. Please contact [redacted]

12 https://www.parliament.uk/business/committees/committees-a-z/commons-select/digital-culture-media-and-sport-committee/news/fake-news-report-published-17-19/

13 See e.g. Avaaz, Correcting the Record: Corrections as an antidote to disinformation (White Paper), 16 April 2020, available: https://secure.avaaz.org/campaign/en/correct_the_record_study/

14 See e.g. ISD Reply All: Inauthenticity and Coordinated Replying in Pro-Chinese Communist Party Twitter Networks, 6 August 2020, available: https://www.isdglobal.org/isd-publications/reply-all-inauthenticity-and-coordinated-replying-in-pro-chinese-communist-party-twitter-networks/; Stanford Internet Observatory Cyber Policy Center, Reporting for Duty: How a Network of Pakistan=Based Accounts Leveraged Mass Reporting to Silence Critics, 1 September 2020, available: https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/20200901_pakistan_report.pdf