CarnegieUK TRUST

CHANGING MINDS • CHANGING LIVES

Submission to the Forum on Information and Democracy Working Group on Infodemics

September 2020

This submission outlines the system-based approach to regulation that Woods and Perrin proposed under the aegis of Carnegie UK Trust to tackle harms on the internet (specifically social media) and how that proposal may operate to mitigate disinformation and misinformation with particular relevance to the four themes of the call for contributions. The response gives the background to the Carnegie UK Trust proposal before looking at the specific issues raised by this call.

a) Background to the work

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 '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 (published 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 government's '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 gave evidence to the International Grand Committee on Fake News; while it did not make specific

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

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]

b) System-based regulation – a statutory duty of care

The work for 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-focussed measures if needed.

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 a regulator with a double-role:

- informing and facilitating good practice (eg through the drafting of codes or guidance); and verification of compliance of the operators with the duty and, where necessary, enforcement.
- 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 – the concept of harm is consistent regardless of the technology and service that might cause it as the 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

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

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.

Response to Consultation

a) Meta Regulation of Content Regulation

Direct content regulation in the context of social media in particular is problematic for a range of reasons, including the amount of data and the speed with which content is uploaded. Assessing content requires an understanding of context which means it is very difficult to develop general rules that apply across different countries. Moreover, the specification of content as acceptable and unacceptable goes to the heart of freedom of expression and raises concerns about the risks of politically motivated suppression of speech. Regulatory obligations which focus on the underlying system, and the extent to which they have been designed with an awareness of the risks of design features and to allow user control, mitigate these concerns. Such an approach could allow the development (within the space allowed by the law) of differently calibrated communities.

So-called 'community standards' are important to meta-regulation. In addition to concerns about the impact of design features on content and information flow, rules focussing on the underlying system should take into account the following aspects. Community standards should be clear, with explanation as to what those standards mean (rather than just legalese or brief statements); the language used should be appropriate for the user group (this is particularly relevant for services that might be used by children, and bearing in mind the different stages of children's development). Community standards should be upfront in platform design (not right at the bottom of page with lots of scrolling ending in click through).

Community standards should be enforced, with visibility as to the enforcement as well as transparency. Requirements should be in place as to resources for user complaints/enforcement of standards (perhaps a percentage of revenue as a benchmark) with requirements to demonstrate provision of resources and relevant training of staff. The Carnegie proposal envisaged that the operation of the complaints and redress mechanisms should be part of an operator's reporting obligations. This should be at a reasonably granular level so that inequality or even discrimination in the system (in terms of whose complaints are taken seriously; the type of complaint that is responded to swiftly) can be identified and tackled.

b) Platform Design

The impact of platform design on users' choices, as well as the flow of information across platforms is central to the duty of care proposal. The connection between design and content can be seen at the following stages, though some issues come up at multiple stages:

- user posting – this concerns sign-in features (the necessity/desirability of user or age verification, but also concerns around whether private groups or encryption are deployed); augmented reality features (e.g. filters and overlays, for example plastic surgery filters); ease or difficulty of embedding content from other platforms; incentives to post clickbait/impact of metrics
- discovery and navigation – this includes recommender algorithms, content curation/personalisation features as well as push notifications; it also includes ex ante moderation (especially that carried out automatically)

- advertisers – does the platform engage any KYC ("know your client") on clients; what ads does it permit? How are audiences segmented (e.g. what controls are there around permitted groupings/ topics – are any segments impermissible or undesirable)
- recipient user – layout of page (what information feeds are prioritised); tools for engaging with content (e.g. like/retweet/upvotes) or controlling/blocking content – are such controls usable and prominent; ability to forward content to individuals or large groups.
- complaints – how easy to use and accessible is the system/ is there a need for appeals does the complaints system work as a reasonable user might expect it to (and taking into account where services are used by children, their different developmental stages)? To what extent is the position of a victim of harm appreciated by the platform bearing in mind different groups may have differing experiences?

The aim of the duty of care approach is not to say that any one design feature is prohibited or mandatory but rather that the platform should assess those features for risk of harm (as set out in the underpinning statute) and to take appropriate steps to mitigate those risks. For example, with regard to encryption perhaps mitigating features relate to identification of users, or limitation of how many users could be in a group using encryption. In terms of advertising, some form of due diligence as to who is using the system and for what ends may be helpful; as well as some form of risk assessment of the categories of audience segmentation. Is there a risk, for example, in segmenting the audience by reference to its members' interest in conspiracy theories/alternative facts, or allowing new parents to be targeted with disproven claims about the link between the MMR vaccine and autism? The availability of micro-targeting itself should be assessed for its risks. The process surrounding risk assessment and mitigation should be documented; the onus should be on the platforms to demonstrate compliance.

In the context of the pandemic, some platforms have sought to emphasise reliable sources of information. While this is a welcome step it is questionable whether this is a sufficient analysis of the issues that seem to lead recommender algorithms and similar features to prioritise (increasingly) extreme or emotive content in general. This response is a piecemeal response, based on individual categories of content rather than a fundamental systems check.

c) Private Messaging

There is a challenge in defining the scope of any such regulatory regime. The Carnegie proposal took the view that the key elements were (1) user generated content; (2) display or making available and (3) multiway communications. The proposal excluded mass media where there was some body with editorial control for content. The proposal also considered whether there should be a threshold below which the regime should not apply. After consideration, the authors rejected this suggestion as some serious harms could occur on small platforms. The proposal was that what was 'reasonable' should take into account the size and resources of the platform.

While some platforms would seem clearly to fall within scope (e.g. Twitter), there are increasing services and tools available for sharing information. This could be a social network platform or a business tools platform (e.g. Slack), video conferencing (e.g Zoom) or even cloud storage (DropBox). Gaming platforms may also raise questions of whether they should within scope, whether concerning communication within the game itself or externally to the game but on game-related platforms. For example, Twitch allows gamers to stream content that the gamers have generated (on games sites) with the intention of interacting with an audience about that content. Twitch provides a place for that display, multiway discussion about it and provides a form of organisation that allows a user to find the particular content

they wish to engage with. Discord, a messaging service for gamers now with over 140 million users, 'almost quadrupled' in size during 2017-18.

While there may be other policy concerns driving decisions about scope, one particular issue deserves focus – that is, issues relating to privacy. Social media in particular have blurred the boundaries between public expression – that is communication to large audiences or that is intended to be viewed by such an audience – and private communication (historically, one-to-one communication typically by letter or telephone conversations). These types of communication have typically been protected from state intrusion, whether in international human rights instruments or national constitutions. This protection should not be defined by reference to the technology alone (so that fact that a service as a matter of domestic law is not a telephony service should not be conclusive as to the scope of privacy). While concerns about state access are less prevalent regarding posts which are open to view generally, questions remain about communication that is closed, that is within a particular group. Yet messaging services and private groups raise real concerns about some of the most serious harms, especially when they are encrypted. Many governments, of course, have communications interception regimes with judicial and/or parliamentary oversight, although the operation of these regimes is frequently contested.

It has become clearer that messaging services have gone beyond one-to-one communications and small groups; many platforms allow large size groups. The size of these groups suggests that the communication mediated via the service is neither private nor confidential. Other characteristics also indicate the non-private nature of the communication, notably the growing practice of public groups, sharing of group links and browsers and search apps for groups. Services that enable the creation of public groups and/or large groups would, in our view, suggest that these services- if they can be used for multiway communications – should fall within a regulatory regime. It seems to us that the only clear boundary is that between one-to-one services (which have traditionally fallen within private life) and one to 'more than one'; the regulatory regime should consider whether the technology may be used to communicate with more than one person not whether it is in a particular instance so used – though potentially this would bring email within regime. Given that many platforms – as well as bad actors – still use email, this may be no bad thing.

While messaging services should be in scope, the Carnegie proposal emphasises, that any risk assessment should take into account the specificities of this form of service. In this context the encrypted nature of a messaging service is a key risk factor; some counter measures could be user identification, limitation on numbers in groups, limitations on how material may be forwarded or searched, limitations on bot membership of groups, or the number of groups which one account may join.

A further question arises with regard messaging services (though also relevant to other platforms as well); that is, the responsibility for aspects provided by third party apps. In particular, third parties have provided apps that allow users to search private groups, to search for membership codes for private groups and the like. Should these apps be the responsibility of the platform provider (this seems unfair save in the instance of lax security which may engage other responsibilities); should the third party apps also be within scope? Conversely where a platform chooses to incorporate a feature using a package from a third party, the service provider should take responsibility for that feature, defects and all.

d) Transparency

Currently, it is hard to assess the effectiveness of platform operators' attempts to ensure safety for their users. Information is not available, and when it is available, it covers matters chosen by the operator in a format chosen by the operator, making assessment and comparison between platforms difficult if not impossible. In the context of the current pandemic a number of social media companies have taken some steps to limit misinformation/disinformation. For example, Pinterest took a decision last year not to have anti-vax material on their platform. Its very clear community guidelines do not allow content that might have "immediate and detrimental" effects to health or public safety, so this allowed them to easily extend it to cover searches for Covid19 and limit the results to material from authoritative sources.[6] WhatsApp's "velocity limiter" reduced the number of times things can be forwarded which, it claims, has led to a 70% reduction in "highly forwarded" messages on its services.[7] There is, however, no way to investigate such a claim. More information is then an essential part of ensuring platforms take responsibility for their products, though on its own it is insufficient. The Carnegie full report suggested the following points that could be required in general – that each relevant operator should be required to:

- develop a statement of assessed risks of harm, prominently displayed to all users when the regime is introduced and thereafter to new users; and when launching new services or features;
  provide its child protection and parental control approach, including age verification, for the regulator's approval;
- develop an internal review system for risk assessment of new services, new tools or significant revision of services prior to their deployment (so that the risk is addressed prior to launch or very risky services do not get launched) – and document this;
- develop a triage process for emergent problems (the detail of the problem may be unknown, but it is fairly certain that new problems will be arising, as the issue of misinformation and disinformation related to Covid19 illustrates) – and document this;
- provide adequate complaints handling systems with independently assessed customer satisfaction targets and also produce a twice yearly report on the breakdown of complaints (subject, satisfaction, numbers, handled by humans, handled in automated method etc.) to a standard set by the regulator, including a self-assessment of performance.

The Carnegie proposal also included information gathering powers for the regulator. As in many other regulatory fields, failure to comply should be a violation of the regime in and of itself.

Further information

We would be happy to provide further information to the working group or discuss further. Please contact maeve.walsh@carnegieuk.org

6 Pinterest's community guidelines say that: "Medically unsupported health claims that risk public health and safety, including the promotion of false cures, anti-vaccination advice, or misinformation about public health or safety emergencies" It also won't have conspiracy theories or content that originates from disinformation campaigns. (See https://policy.pinterest.com/en-gb/community-guidelines)

7 https://techcrunch.com/2020/04/27/whatsapps-new-limit-cuts-virality-of-highly-forwarded-messages-by-70/?guccounter=1