CarnegieUK TRUST

CHANGING MINDS • CHANGING LIVES

Submission to UN High-Level Panel on Digital Co-operation

January 2019

This response from William Perrin and Professor Lorna Woods puts forward, as a response under section III of the Panel's inquiry (Illustrative Action Areas), a proposal for harm reduction on social media. Developed for Carnegie UK Trust, the authors believe that this approach can address challenges arising in a number of the Panel's identified areas, particularly the protection of human rights online and digital trust and security. William Perrin presented these proposals at the meeting of the UN High-Level Panel on Digital Co-operation symposium in London on 10 December 2018; this response provides more detail and supporting references, as well as an update on some of the thinking set out in a recently published paper by William Perrin and Professor Woods.

III a) What are the challenges faced by stakeholders (e.g. individuals, Governments, the private sector, civil society, international organizations, the technical and academic communities) in these areas?

1. In the UK during the past year, there has been increasing acknowledgement amongst policymakers and politicians of the extent and impact of the harms caused to users of online services. For the past 18 months, the UK government has been developing and consulting on its Internet Safety Strategy White Paper – now expected to be published in spring 2019 as a joint publication from the Department for Digital, Culture, Media and Sport (DCMS) and the Home Office. In May 2018, the then Secretary of State for DCMS confirmed that the government was intending to legislate to address online harms and that the forthcoming White Paper would:

   > *set out plans for upcoming legislation that will cover the full range of online harms, including both harmful and illegal content. Potential areas where the Government will legislate include the social media code of practice, transparency reporting and online advertising.'[1]*

2. Since then, a proliferation of Parliamentary inquiries [2] have been taking evidence from stakeholders across industry, civic society and academics on various aspects of online harms, while many research and campaigning organisations have put forward their own proposals for regulation. [3] More are expected in the run-up to the government's White Paper.

3. Evidence of harms is also building but is still partial. A recent OFCOM/ICO research paper [4] 'Internet users' experience of harm online' has helpfully established an evidence base of harms independent of lobby groups. OFCOM's sample size of 1,600 is an order of magnitude better than

most extant research, this is short of large-scale multi-annual randomised control trials, which are very difficult to secure in an area of innovative technology that suffers from waves of fashion in its user base. In academic literature, there is a lot of emerging evidence of correlation between some forms of online activity and harms to individuals [5], but research in this area struggles to identify causation. The UK's Chief Medical Officer is expected to publish her view on the available evidence of social media harms to young people's health in late Spring. [6]

4. As a result, rapidly-propagating social media services, subject to waves of fashion amongst young people, are a particular challenge for legislators and regulators. The harms are multiple, and may be context- or platform- specific, while the speed of their proliferation makes it difficult for policymakers to amass the usual standard of long-term objective evidence to support the case for regulatory interventions.

5. That is why, in developing our proposal for Carnegie over the past year [7], we have looked to other regulatory regimes for a workable, principle-based approach to reduce the risk of harms to individuals. Our updated paper provides some new thinking and refinement of our initial proposals following feedback and consultation with stakeholders. [8]

6. We set out in the next section how the application of a statutory "duty of care" to social media and online services would work in practice, and the roles of the main players in developing and implementing it.

The precautionary principle

7. The traditional approach of not regulating innovative technologies needs to be balanced with acting where there is good evidence of harm but there has not been enough time to establish authoritative evidence. We see this as a core challenge for establishing and operating a new regulatory regime and have sought a robust basis for action in the face of scientific uncertainty.

8. After the many public health and science controversies of the 1990s, the UK government's Interdepartmental Liaison Group on Risk Assessment (ILGRA) published a fully worked-up version of the precautionary principle for UK decision makers. [9]

> *'The precautionary principle should be applied when, on the basis of the best scientific advice available in the time-frame for decision-making: there is good reason to believe that harmful effects may occur to human, animal or plant health, or to the environment; and the level of scientific uncertainty about the consequences or likelihoods is such that risk cannot be assessed with sufficient confidence to inform decision-making.'*

9. The ILGRA document advises regulators on how to act when early evidence of harm to the public is apparent, but before unequivocal scientific advice has had time to emerge, with a particular focus on novel harms. The ILGRA's work is still current and hosted by the Health and Safety Executive (HSE), underpinning risk-based regulation of the sort we propose.

A duty of care

10. Social media platforms are forms of public spaces. People go to such platforms for all sorts of activities and, while using them, should be protected from reasonably foreseeable harm as they would expect in any public place, such as an office, bar or theme park. Owners of public places in the physical realm are bound by a duty of care [10], a concept straightforward in principle and well-established, such as in the Health and Safety at Work Act 1974. A person (including companies) under a duty of care must take care in relation to a particular activity as it affects particular people or things. If that person does not take care and someone comes to harm as a result then there are legal consequences, primarily through a regulatory scheme but also with the option of personal legal redress.

11. Applying the "duty of care" approach to the social media sphere has a number of significant benefits:

    - It is simple, broadly-based and largely future-proof – expressed in terms of outcome (the prevention of harm) not specifics of process. [11]

    - The regulatory approach is essentially preventative, reducing adverse impact on users before it happens, rather than a system aimed at compensation/redress.

    - The categories of harm can be specified at a high level, by Parliament, in statute. [12]

    - It would apply to all social media service providers regardless of size, with the regulator taking a proportionate approach according to the severity of harm and the size of risk, as well as the size of, and resources available to, a service operator.

    - A risk-based regulatory approach provides for safe system design (including operational and business choices), which we believe should be backed by a statutory safety by design code.

    - In micro economic terms, it returns external costs to the production decision and is efficient if applied in a manner proportionate to risk of harm.

III b) What are successful examples of cooperation among stakeholders in these areas? Where is further cooperation needed?

12. The emerging consensus in the UK is that self-regulation of social media has failed [13]; the exact nature of regulation proposed by the government remains to be seen but accountability from social media and online service providers for the harms caused by their services is long overdue.

13. Since publishing our original proposals, we have revisited Lawrence Lessig's work from 1999 [14]. Lessig observed that computer code sets the conditions on which the internet (and all computers)

is used. While there are other constraints on behaviour (law, market, social norms), code is the architecture of cyberspace and affects what people do online: code permits, facilitates and sometimes prohibits. It is becoming increasingly apparent that it also nudges us towards certain behaviour. While Lessig's work was oriented along a different line, it reminds us that the environment within which harm occurs is defined by code that the service providers have actively chosen to deploy, their terms of service or contract with the user and the resources they deploy to enforce that. Service providers could choose not to deploy risky services without safeguards or they could develop effective tools to influence risk of harm if they choose to deploy them.

14. In sum, online environments reflect choices made by the people who create and manage them; those who make choices should be responsible for the reasonable foreseeable risks of those choices.

15. Our latest thinking also looks at the responsibilities for harm reduction beyond the immediate service or platform. This reflects evidence that:

    a) harm may implicate more than one platform and, more generally,

    b) people are harmed by content on social media services even when they themselves are not customers of those services.

16. As regards (a) we note that there are some examples of good practice. For example, Twitch is already grappling with this third-party service problem. After user feedback, Twitch gave itself powers [15] to sanction customers who use another service (Twitter, say) to organise attacks on a fellow Twitch user. Twitch extends this to IRL meetups. Twitch requires evidence to be presented to it. This suggests that a provider's responsibility does not end with the limits of its own platform and that deliberate offenders will move from platform to platform. We note that the process of regulation could bring service providers of all types together to share knowledge about harms within and between platforms, putting commercial interests to one side. [16]

17. As regards (b): consider the harm suffered by a woman who has revenge porn posted on a service of which she is not a customer. The service provider's obligation to the victim should not depend on whether or not she had signed up to the service that was used to harass her. Extending the statutory duty to individuals who are not users of the service is important as it is far from certain that, under the common law duty of care, a duty would arise to such an individual.

III c) What form might cooperation among stakeholders in these areas take? What values and principles should underpin it?

18. While necessarily underpinned by regulation, we envisage the implementation of a statutory duty of care to be a co-operative and iterative process with multiple stakeholders involved in the "harm reduction cycle".

The harm reduction cycle

19. New legislation would set out the duty of care and identify the key harms Parliament wants the regulator to focus on. We suggest that those harms would be: the 'stirring up of hatred offences', national security, harms to children, emotional harm, harms to the judicial and electoral processes, economic harms, though work needs to be done on the scope of each of the harm.

20. We suggest that the regulator runs a harm reduction cycle as set out in the diagram (overleaf), involving civil society as well as companies at each consultative step. The regulator would begin by requiring companies to measure and survey harm: the regulator would draw up a template for measuring harms, covering scope, quantity and impact. The regulator would use as a minimum the harms set out in statute but, where appropriate, include other harms revealed by research, advocacy from civil society, the qualifying social media service providers etc. The regulator would then consult publicly on this template, specifically including the qualifying social media service providers.

21. The qualifying social media services would then run a measurement of harm based on that template, making reasonable adjustments to adapt it to the circumstances of each service. The regulator would have powers in law to require the qualifying companies (see enforcement below) to comply. The companies would be required to publish the survey results in a timely manner. This would establish a first baseline of harm.

22. The companies would then be required to act to reduce these harms. We expect those actions to be in two groups – things companies just do or stop doing, immediately; and actions that would take more time (for instance new code or terms and conditions changes). Companies should seek views from users as the victims of harms or NGOs that speak for them. These comments – or more specifically the qualifying social media service providers respective responses to them (though it should be emphasised that companies need not adopt every such suggestion made) – would form part of any assessment of whether an operator was taking reasonable steps and satisfying its duty of care. Companies would be required to publish, in a format set out by the regulator:

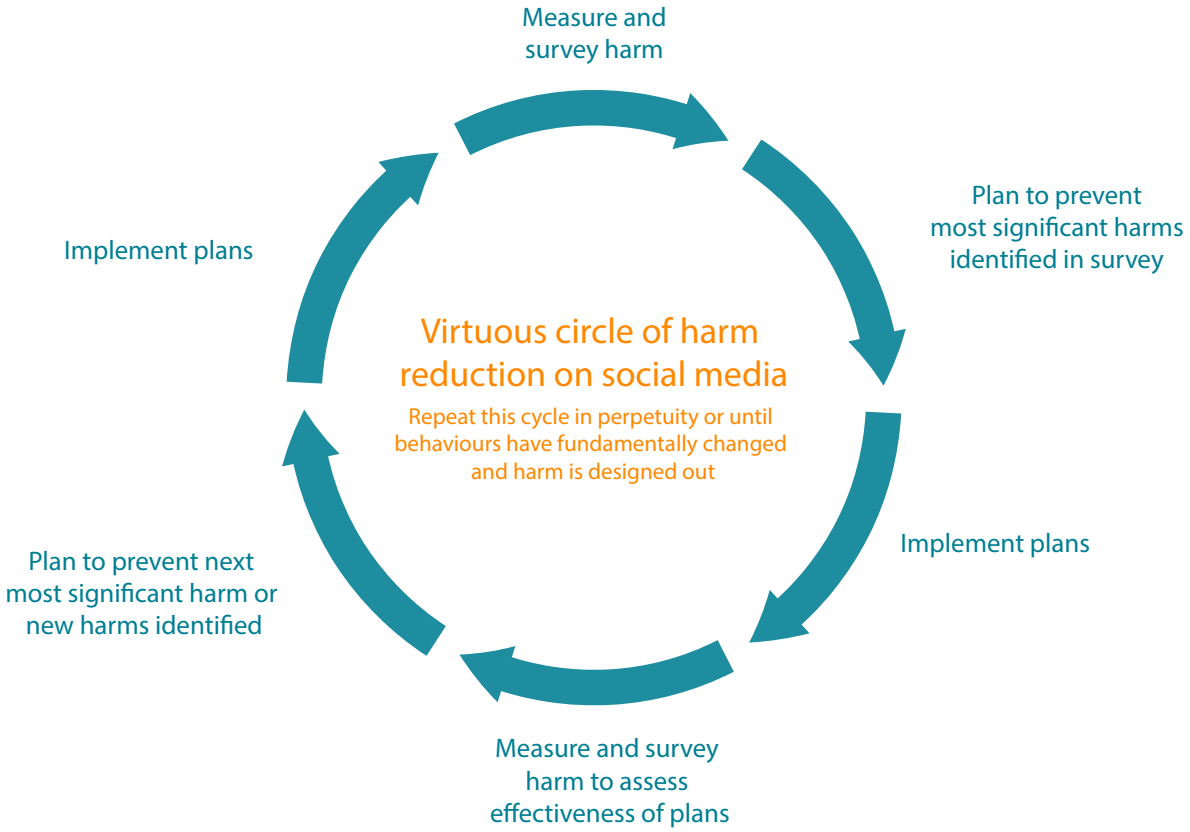
    a) what actions they have taken immediately;

    b) actions they plan to take;

    c) an estimated timescale for measurable effect; and

    d) basic forecasts for the impact on the harms revealed in the baseline survey and any others they have identified.

23. The regulator would invite views on the plan from the public, industry, consumers/users and civil society and make comments on the plan to the company, including comments as to whether the plan was sufficient and/or appropriate. The companies would then continue or begin their harm reduction work based on their individual plans.



24. Harms would be measured again after a sufficient time has passed for harm reduction measures to have taken effect, repeating the initial process. This establishes the first progress baseline. We have recently taken the view that an output of this cycle would be codes of practice that could be endorsed by the regulator. In our view, the speed with which the industry moves would mitigate against traditional statutory codes of practice which require lengthy consultation cycles. The government, in setting up such a regime, should allow some lee-way from standard formalised consultation and response processes.

25. However, if harms surveyed in the baseline have risen or stayed the same, the companies concerned will be required to act and plan again, taking due account of the views of victims, NGOS and the regulator.

26. In these instances, the regulator may take the view that the duty of care is not being satisfied and, ultimately, may take enforcement action. The regulator would set an interval before the next wave of evaluation and reporting. If the cycle does not reduce harms or the companies do not co-operate then sanctions could be deployed. These might include, for example:

- Administrative fines in line with the parameters established through the Data Protection Bill regime of up to €20 million, or 4% annual global turnover – whichever is higher.

- Enforcement notices – (as used in data protection, health and safety) – in extreme circumstances a notice to a company to stop it doing something. Breach of an enforcement service could lead to substantial fines.

- Enforceable undertakings where the companies agree to do something to reduce harm.

- Adverse publicity orders – the company is required to display a message on its screen most visible to all users detailing its offence

- Forms of restorative justice – where victims sit down with company directors and tell their stories face to face.

27. We have been doing further thinking on how to ensure a sanctions and penalties regime bites and our most up-to-date ideas on corporate and director responsibility are set out in our recent paper.

Next steps

28. While we await the proposals from the UK government, we continue to discuss the merits of a statutory "duty of care" with stakeholders, regulators and policymakers and to refine and expand the ideas. Even if the DCMS and Home Office White Paper sets out a commitment to regulation to reduce online harms, the timescales for policy development, consultation and the introduction and debate of subsequent legislative proposals in Parliament means that its implementation would be a long way off. Meanwhile, evidence of harms continues to increase. Grounded in established statutory approaches from other regimes and requiring – we think – a very short Bill to enact some initial regulations, we hope that the "duty of care" approach can be adopted as a first, necessary and urgent step to address some of the most serious harms ahead of the implementation of the whole regime.

William Perrin

Professor Lorna Woods

References

[1] Government Response to the Internet Safety Strategy Green Paper (DCMS, May 2018) : https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/708873/Government_Response_to_the_Internet_Safety_Strategy_Green_Paper_-_Final.pdf

[2] DCMS Select Committee: "Disinformation and Fake News"; Lords Communications Select Committee: "The Internet: To Regulate or Not to Regulate?"; Science and Technology Committee: "The impact of social media on young people's mental health"

[3] For example: https://www.nspcc.org.uk/what-we-do/campaigns/wild-west-web/; https://doteveryone.org.uk/project/regulating-for-responsible-technology/; https://institute.global/insight/renewing-centre/tony-blairs-foreword-new-deal-big-tech; https://webrootsdemocracy.org/kinder-gentler-politics/

[4] See https://www.ofcom.org.uk/research-and-data/internet-and-on-demand-research/internet-use-and-attitudes/internet-users-experience-of-harm-online

[5] https://www.nhs.uk/news/mental-health/worrying-rise-reports-self-harm-among-teenage-girls-uk/#where-does-the-study-come-from; ISER study: https://doi.org/10.1186/s12889-018-5220-4

[6] The Chief Medical Officer has been tasked to survey evidence on social media: https://www.theguardian.com/politics/2018/apr/22/jeremy-hunt-social-media-firms-failing-safeguard-children-online

[7] Blog Posts: https://www.carnegieuktrust.org.uk/project/harm-reduction-in-social-media/ ; Evidence to Lords Communications Select Committee: http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/communications-committee/the-internet-to-regulate-or-not-to-regulate/written/82684.html

[8] https://www.carnegieuktrust.org.uk/publications/internet-harm-reduction/

[9] http://www.hse.gov.uk/aboutus/meetings/committees/ilgra/pppa.htm

[10] Occupiers Liability Act 1957 S2 http://www.legislation.gov.uk/ukpga/Eliz2/5-6/31/section/2

[11] The UK government recently confirmed that the 1974 duty of care in the Health and Safety at Work Act applies to artificial intelligence software employed in the workplace. https://www.parliament.uk/business/publications/written-questions-answers-statements/written-question/Lords/2018-05-23/HL8200/

[12] See Health and Safety at Work Act 1974 S2: http://www.legislation.gov.uk/ukpga/1974/37/section/2

[13] Petitions Committee Report on Online Abuse and Disabled People (January 2019): https://www.parliament.uk/business/committees/committees-a-z/commons-select/petitions-committee/news-parliament-2017/oa-report-publication-22-01/

[14] See Lawrence Lessig, "The Law of the Horse: What Cyberlaw Might Teach", (1999), 113 Harv. L. Rev. 501; also Code and Other Laws of Cyberspace (1999) and Code: version 2.0 (2006)

[15] Twitch corporate blog announcing changes (8 February 2018) https://blog.twitch.tv/twitch-community-guidelines-updates-f2e82d87ae58; 'We may take action against persons for hateful conduct or harassment that occurs off Twitch services that is directed at Twitch users.' (Twitch Community Guidelines on Harassment: https://www.twitch.tv/p/legal/community-guidelines/harassment/)

[16] As the service providers do to counter terrorism in the Global Internet Forum to Counter Terrorism: https://gifct.org/about/