CarnegieUK
TRUST
CHANGING MINDS • CHANGING LIVES

RESPONSE TO COMMITTEE ON STANDARDS IN PUBLIC LIFE REVIEW OF ARTIFICIAL INTELLIGENCE AND PUBLIC STANDARDS

July 2019

1. This short note makes a primary factual observation and covers a full reference paper that sets out work we have carried out to develop a proposal for a statutory duty of care for harm reduction on social media. We would commend this work to the Committee as they start their deliberations for this important review.

AI safeguards and the accountability principle of public life – where are the Health and Safety at Work Act 1974 risk assessments?

2. William Perrin, Trustee of Carnegie UK Trust, worked with Lord Stevenson of Balmacara to establish whether the provision of the Health and Safety at Work Act (HSAWA) 1974 applied to artificial intelligence systems used in decision support that may be 'articles supplied for use at work' in Section 6 of the Act. The Government is adamant that it does. And that this 'requires such testing and examination as may be necessary to ensure that any article for use at work is safe and without risks'. The exchange with Baroness Buscombe is below. In our work on online harms we have found few who understand that this existing regulatory regime applies to AI/ML etc.

3. The Committee might wish to follow up on the spirit of the accountability principle by asking government bodies for risk assessments and testing they have had carried out on AI systems and discuss the matter with the Health and Safety Executive. More broadly the Committee might wish to consider how the statutory duties of care in the HSAWA 1974 are being implemented by the government when they deploy AI and ML decision support systems.

   Lord Stevenson of Balmacara 23 May 2018 to Department for Work and Pensions[1]

   Industrial Health and Safety: Artificial Intelligence HL8200

   To ask Her Majesty's Government what assessment they have made of the extent to which section 6 of the Health and Safety at Work etc. Act 1974 applies to artificial intelligence or machine learning software that is used in the workplace to (1) control or animate physical things in the workplace, (2) design articles for use in the workplace, or (3) support human decision-making processes running on computers under the control of the employer with an impact on people's health and safety; and whether, in each case, testing regimes exist as set out in section 6(1)(b) of that Act.

   Answered by: Baroness Buscombe 05 June 2018

[1] https://www.parliament.uk/business/publications/written-questions-answers-statements/written-question/Lords/2018-05-23/HL8200/

Section 6 of the Health and safety at Work etc. Act 1974 places duties on any person who designs, manufacturers, imports or supplies any article for use at work to ensure that it will be safe and without risks to health, which applies to artificial intelligence and machine learning software. Section 6(1)(b) requires such testing and examination as may be necessary to ensure that any article for use at work is safe and without risks but does not specify specific testing regimes. It is for the designer, manufacturer, importer or supplier to develop tests that are sufficient to demonstrate that their product is safe.

The Health and Safety Executive's (HSE) Foresight Centre monitors developments in artificial intelligence to identify potential health and safety implications for the workplace over the next decade. The Centre reports that there are likely to be increasing numbers of automated systems in the workplace, including robots and artificial intelligence. HSE will continue to monitor the technology as it develops and will respond appropriately on the basis of risk.

A statutory duty of care for social media harm reduction

4. In 2018-2019, Professor Lorna Woods (Professor of Internet Law in the School of Law at the University of Essex) and William Perrin (a Carnegie UK Trustee and former UK government Civil Servant) developed a public policy proposal to improve the safety of some users of internet services in the United Kingdom through a statutory duty of care enforced by a regulator. Woods and Perrin's work under the aegis of Carnegie UK Trust took the form of many blog posts, presentations and seminars.

5. A full reference paper drawing together their work on a statutory duty of care was published in April 2019, just prior to the publication of the Online Harms White Paper. We have attached it as an annex to this paper and it can be viewed, along with all the other material relating to this proposal and a full recent response to the DCMS consultation on the Online Harms White Paper, on the Carnegie UK Trust website: https://www.carnegieuktrust.org.uk/project/harm-reduction-in-social-media/

6. Our work has influenced the recommendations of a number of bodies, including: the House of Commons Science and Technology Committee, the Lords Communications Committee, the NSPCC, the Children's Commissioner, the UK Chief Medical Officers, the APPG on Social Media and Young People and the Labour Party.[2] A statutory duty of care has been adopted – though not fully as we envisaged – by the Government as the basis for its Online Harms White Paper proposals[3]. Most recently, though it did not refer to our work, a report to the French Ministry of Digital Affairs referenced a "duty of care" as the proposed basis for social media regulation.[4]

7. We urge the Committee and its review team to read our reference paper in full. Any discussion of how to set standards for AI use in any sector will need to take account of a wide range of technical

[2] 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://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

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

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

and ethical issues. But we believe that taking a step back may help consider the issues in a simpler way: our proposition for a systemic duty of care to reduce reasonably foreseeable harms occurring on social media – based on well-established frameworks in areas such as health and safety – is a good starting point for all considerations of how to ensure the safe, ethical and fair deployment of any technology that has an impact on end users.

8. There are three particular areas where we believe our work has relevance for any considerations around the deployment of AI and the impact on public life:

   i) AI as the algorithm that promotes content. From a duty of care perspective, we would see this as driving the sorts of content that is promoted and the sorts of content that is excluded from searches/autoplay, all of which are design choices made by the platform or service provider, and within this we would also consider tools that are available from third parties eg that allow a user to search hashtags;

   ii) the role of AI in creating content – eg the technology behind deepfakes; while this is not directly covered in our work, it is flagged in the DCMS Online Harms White Paper as a contributing factor to harms caused by disinformation, by which it is becoming even easier to create and disseminate false content and narratives (p23)

   iii) the role of AI in spotting the problem, eg identifying and removing illegal or harmful content, where we see a number of particular concerns:

   - the focus on this makes it an ex post issue when we would see a duty of care requiring companies to consider more basic questions of platform design;

   - the material that is being used for training: there are possible problems with an unequal coverage (eg language but also bias);

   - does the focus on AI distract from other questions (eg focus on design choices)?

9. There is also an intersection between the Committee's inquiry and the Online Harms White Paper in relation to emergence of harms to democracy. Along with other civic society organisations, we have published a statement calling for a greater focus on societal harms, such as those which impact democracy, in the Government's White Paper proposals and the scope of harms to come under the regulatory regime.

10. Our work draws on a number of established legal and policymaking frameworks but we would particularly draw the Committee's attention to our discussion, in chapter two of the attached paper, on the application of the precautionary principle to the question of how to regulate or set standards for innovative, fast-developing technologies:

    The government has often been called to act robustly on possible threats to public health before scientific certainty has been reached. 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: '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.'* [5]

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. ILGRA's work focuses on allowing economic activity that might be harmful to proceed 'at risk', rather than a more simplistic, but often short-term politically attractive approach of prohibition. 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.

11. We would also particularly refer the Committee to our discussion, in chapter 3, on the design decisions that drive any of our interactions with technology – whether viewing content online, or engaging with AI-enabled services. These environments are defined by code that the service providers have actively chosen to deploy, their terms of service or contract with the user and the resources service providers deploy to enforce that. While technological tools can be used for positive reasons as well as have negative impacts, it is important to remember that they are not neutral,[6] nor are they immutable. Corporate decisions drive what content is displayed to a user. 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.

12. These decisions are best taken when informed by a risk assessment. There will be risks which will be obvious – for instance material harm is known to have occurred before in certain circumstances and those which, while not obvious are foreseeable. If a material risk is foreseeable then a company should take reasonable steps to prevent it. This is as true – and even more important – in relation to the deployment of AI as it is in relation to any other technology or service.

13. There are many moving parts in this landscape, and many government and regulatory organisations undertaking concurrent reviews of bits of it. Protecting users from harm – however it manifests itself - has to be at the heart of all those proposals. Given the commitment of the Government to introduce a statutory duty of care to reduce online harms, we would urge the Committee on Standards in Public Life to consider how their review might take account of its principles and consider how the thinking that underpinned our work can apply to theirs.

14. We are happy to speak to you further about our proposals or assist in any way in the Committee's review.

Carnegie UK Trust
July 2019

[5] 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

[6] W. Hartzog, Privacy's Blueprint: The Battle to Control the Design of New Technologies (Cambridge, MA: Harvard University Press, 2018)