Written evidence submitted by Trustworthy Autonomous Systems Hub (OSB0060)

Authors

Professor Sarvapali D. Ramchurn: Professor of AI and Director of the UKRI Trustworthy Autonomous Systems (TAS) Hub. Sarvapali has over 19 years of experience in developing AI solutions.
Dr Stuart E. Middleton: Lecturer in Computer Science, University of Southampton. Middleton has over 17 years of experience researching natural language processing in projects involving law enforcement agencies, criminal behaviour and online harms. He was an invited expert at the 2019 UK Cabinet Office ministerial AI roundtable on 'use of AI in policing'.
Professor Derek McAuley: Professor of Digital Economy at the University of Nottingham, Director of Horizon Digital Economy Research Institute, a Research Hub within the UKRI Digital Economy programme, and Deputy Director of the UKRI TAS Hub.
Dr Helena Webb: Transitional Assistant Professor, Horizon Digital Economy Research Institute, University of Nottingham. She is a highly experienced socio-technical researcher who has worked on projects including the UKRI funded 'Digital Wildfire' study, which investigated the spread of harmful content on social media and opportunities for the responsible governance of digital social spaces, and UnBias explored the user experience of algorithm-driven online platforms.
Dr Richard Hyde Professor of Law, Regulation and Governance at the University of Nottingham. Professor Hyde is a non-practicing solicitor with a particular interest in the regulation of novel technologies.
Dr Justyna Lisinska: She is a Policy Research Fellow at King's College London. Justyna completed a PhD degree in Web Science at the University of Southampton. Her thesis investigated online political discussion on Facebook pages of populist supporters. She has experience of working at the Cabinet Office, providing policy recommendations.

About the TAS Hub:

The UKRI TAS Hub assembles a team from the Universities of Southampton, Nottingham and King's College London. The Hub sits at the centre of the £33M Trustworthy Autonomous Systems Programme, funded by the UKRI Strategic Priorities Fund. The role of the TAS Hub is to coordinate and work with six research nodes to establish a collaborative platform for the UK to enable the development of socially beneficial autonomous systems that are both trustworthy in principle and trusted in practice by individuals, society and government. Read more about the TAS Hub here.

Citation:
Trustworthy Autonomous Systems Hub et al. (2021) A Response to Draft Online Safety Bill - a call for evidence from the Joint Committee. DOI: https://doi.org/10.18742/pub01-060

Content in scope

1. ***Are the definitions in the draft Bill suitable for service providers to accurately identify and reduce the presence of legal but harmful content, whilst preserving the presence of legitimate content?***

We consider the current draft of the online safety bill does not adequately protect children (or vulnerable adults) from the harmful impact of aggregated low-level cyberbullying content, which when consumed by children over a long period of time has the potential to present a serious risk of psychological impact and potential for harm (e.g. depression, self-harm or suicide). When considered individually, low-level cyberbullying content items will often not meet the threshold to be considered harmful by a service provider. However, when content is considered in bulk, such as niche forum groups promoting suicide, or over a long period of time, such as continuous cyberbullying over a year, the risk of harm should meet the threshold increases to be considered harmful by a service provider.

It is becoming well established that there is a correlation, and likely causal link, between online cyberbullying and harmful mental health outcomes such as depression, self-harm and suicidal ideas [Brailovskaia et al. 2018]. A recent study by Google [Thomas et al. 2021] has mapped out the modern landscape of online abuse, and this includes low-level toxic content such as bullying, trolling, purposeful embarrassment and sexual harassment. Other recent work includes a study of online microaggression towards female members of Parliament [Harmer and Southern, 2021], research workshops [Middleton et al. 2020] and research projects [SafeSpacesNLP, 2021] [ProTechThem, 2021] investigating AI methods to support the classification of online harm behaviour patterns.

Other instances of legal but harmful content that can present risks to children over a period of time in addition to cyber bullying are: content promoting disordered eating behaviours; easy access to sexualised and violent material designed for adult age groups, content advocating/celebrating behaviours of self-harm, content promoting unrealistic idealised body shapes - in particular through edited images which are not marked as such [Internet Matters, 2021; NSPCC Learning, 2021].

RECOMMENDATION #1
We recommend adding an additional category (h) to the list in [part 2, section 7 (9)] around how a 'children's risk assessment' is defined:

> "(h) the level of risk of harm to children presented by exposure to an aggregated bulk of low- level harmful content over an extended time period or within closed communities or groups."

We recommend also adding the same additional category (h) to the list in [part 2, section 7 (10)] around how an 'adult's risk assessment' is defined.

As a result of clause 62(1) OFCOM will be required to provide guidance on what amounts to "low-level harmful content" and "an extended time period." Examples of low-level harmful content that users are exposed to over an extended period of time are cited above.

2. ***Earlier proposals included content such as misinformation/disinformation that could lead to societal harm in scope of the Bill. These types of content have since been removed. What do you think of this decision?***

Academic work has long argued that the reasons behind the generation of misinformation are highly social. For instance, the sociology of rumour indicates that misinformation emerges at items of societal unease and tension, filling gaps where uncertainties exist and allowing groups of people to make sense of the world around them [e.g. Shibutani, 1966; Kapferer 1999]. However, the affordances of online platforms enable this misinformation, once generated, to spread much more rapidly than in other formats. The wide reach of online sites and the sharing functions on many social media platforms mean that pieces of (mis)information can reach very large numbers of people in a very short period of time and become well established [Webb et al. 2016]. A 2013 report by the World Economic Forum [WEF 2013] notes that the speed of the spread of online misinformation can outstrip the capacity for other agencies (traditional news media, police, government etc.) to respond to it in real-time, meaning that it can be difficult to prevent people believing and acting on it. In addition, there is a growth of 'clickbait factories' set up to generate misleading content (for monetary and/or political gain) and push it out online over a short period of time via bot accounts etc. These take advantage of the trending mechanisms of social media platforms to increase the visibility of such misinformation. [Lewandowsky et al. 2017].

Despite these indications, it is very difficult to directly assess the capacity of online misinformation to cause social harm and measure its influence on voting behaviour, health behaviour etc. Furthermore, suggestions to deal with the issue by simply removing potentially misleading content raise concerns about the fundamental right to freedom of expression. The same WEF report mentioned above describes the crucial importance of freedom of speech in democratic societies and the opportunities that online platforms provide for citizens to express their opinions and whistleblow important truths anonymously. Requiring online platforms to delete content, would require social media and internet companies to decide what counts as misleading content and what type of misleading content leads to social harm. However, these decisions might be highly subjective due to the lack of clear criteria of what counts as misleading content [e.g. Bernal, 2017; Whon et. al. 2017]. It is also difficult to know what factors companies took into consideration while dealing with misinformation since this kind of information may be kept confidential.

However, research does show an increasing use by the general public of online sources for news. Survey research conducted by the Reuters Institute for the Study of Journalism [Newman et al. 2017] found that 51% of people with online access use social media as a news source. Whilst for many this use is in addition to traditional news media, in particular television, 28% of 18-24 year olds said that social media was their main news source. This increased use by itself indicates a need to take the prevalence of online misinformation very

seriously. In addition, if this increased use is accompanied - as theorised - by a decline of public trust in traditional news and authority sources [Lewandowsky et al. 2017], plus a loosening of previously established boundaries between (fact-checked, sourced etc.) journalistic content and lay-generated content [Webb and Jirotka, 2017], then the risks of online misinformation having detrimental impacts increase substantially.

RECOMMENDATION #2

Online misinformation is a potentially very serious problem. However, it is also deeply complex and bound up with many social factors and questions of freedom of expression that need careful consideration. We think that it was a good decision to remove misinformation/disinformation from the scope of the Bill. We recommend that the government only include misinformation/disinformation in scope of the Bill if it can provide a clear definition of what counts as misleading content leading to societal harm and develop robust ways of monitoring what content is removed by companies to protect freedom of expression.

Algorithms and user agency

1. ***What role do algorithms currently play in influencing the presence of certain types of content online and how it is disseminated? What role might they play in reducing the presence of illegal and/or harmful content?***

The use of artificial intelligence (AI) algorithms offers service providers an ability to scale up content analysis and perform filtering and triage of user-generated content prior to human moderation. It must be noted, however, that AI models often exhibit the bias that pre-exists in the training data they are trained upon. Modern AI algorithms use deep learning techniques ,which require very large numbers of training examples (100,000+), and these very large training sets are sourced primarily from existing research community resources such as large image datasets, compiled and annotated in over many years, or web-scale text corpora, derived from sources such as Wikipedia. Such datasets [Bolukbasi et al. 2016] contain historical bias (e.g. old fashioned opinions written in 20-year-old news articles), socioeconomic and demographic bias (e.g. face detection datasets consisting of mostly white American males) and bias due to data sparsity (e.g. geographic datasets missing information on global south countries).

Bias in AI models is usually identified by running a set of test examples through a model and checking the results. When AI models are expected to process new types of test data, of types unseen in the training set, it should be expected that new bias and error characteristics may emerge. Major changes in the use of AI models should trigger an updated risk assessment by service providers using them, to allow a change for re-assessment of the models with regards to their propensity for bias and error in light of the new content they are being asked to process.

RECOMMENDATION #3

The draft Online Safety Bill defines in [part 2, section 7 (1)(c)] a duty on service providers to carry out a further illegal content risk assessment before making any significant change to any aspect of the design or operation of a service. For the avoidance of doubt regarding the scope of "operation" we recommend this duty is extended to cover significant changes in (a) the type of content being processed and/or (b) the scale of content being processed. Both changes in type and scale of content are potential triggers for new bias behaviour in AI models used by services, and an updated risk assessment should be conducted to explicitly evaluate this risk.

We recommend adding a new category (d) to the list in [part 2, section 7 (1)]:
"(d) to carry out a further illegal content risk assessment before making any significant change to the type or volume of content processed by a service"

As a result of clause 62(1) OFCOM will be required to provide guidance on what amounts to a significant change in the type or volume of content processed by a service.

The role of Ofcom

1. ***Are the media literacy duties given to Ofcom in the draft Bill sufficient?***

Section (3) of Chapter 8 states that OFCOM *must, in particular, carry out, commission or encourage educational initiatives designed to improve the media literacy of members of the public*. We welcome educational initiatives, but we want to flag the different needs of different "members of the public". For example, educational initiatives implemented from the early years at schools can possibly improve media literacy, but we cannot use the same approach for older generations or vulnerable users who might have different requirements. What is more, the problem of information literacy is not limited to one age group (e.g. digital natives - people who have grown up with modern technologies). For example, a restaurant in London heard accusations from people who believed in news found on a prank website [BBC NEWS 2017].

RECOMMENDATION #4
We recommend the OFCOM considers different initiatives (perhaps not only educational) in the context of socio-demographic differences and the diverse needs of public members. We also recommend a change of the drafting of the clause. OFCOM currently has limited possibilities and can only do what is permitted by the statute. OFCOM should be able to pursue all initiatives (not only educational) designed to improve the media literacy of members of the public.

We also strongly encourage that OFCOM works with academia to understand better media literacy competencies and requirements of various groups of online users. OFCOM should investigate possible interventions by public service broadcasters, whether through traditional broadcast media or their online presence, to reach the widest audience with engaging as well as educational content. However, OFCOM should not only measure the success of campaigns presented to the public via public service broadcasters based on how many people viewed or liked the content. Some other methods for measuring the increase in media literacy/competencies would need to be developed.

References

BBC NEWS (2017) *Restaurant hit by human meat fake news claims*. Available from: http://www.bbc.co.uk/newsbeat/article/39966215/restaurant-hit-by-human-meat-fake-news-claims.

Bernal, P. (2017) *Fake news: if you care about being lied to you'll be more careful about the way you use social media*. Available from: https://theconversation.com/fake-news-if-you-care-about-being-lied-to-youll-be-more-careful-about-the-way-you-use-social-media-77431?utm_campaign=Echobox&utm_medium=Social&utm_source=Facebook#link_time=1494947760

Bolukbasi, T. Chang, K. Zou, J. Saligrama, V. Kalai, A. (2016) Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems* (NIPS'16). Curran Associates Inc., Red Hook, NY, USA, 4356–4364.

Brailovskaia, J. Teismann, T. Margraf, J. (2018) Cyberbullying, positive mental health and suicide ideation/behavior, *Psychiatry Research*, Volume 267.
.

Harmer, E. Southern, R. (2021) Digital microaggressions and everyday othering: an analysis of tweets sent to women members of Parliament in the UK*, Information, Communication & Society.* DOI: https://doi.org/10.1080/1369118X.2021.1962941

Internet Matters (2021) *Social media: facts and advice*. Available from: https://www.internetmatters.org/resources/social-media-advice-hub/?gclid=CjwKCAjw7fuJBhBdEiwA2lLMYUpMgQE8t8zM2Gcfrt9mUwcvEraSz3Bgnzi6Cjzhx7yE0P5DM4sd1xoCoe4QAvD_BwE

Kapferer, J-N. (1990) *Rumors: Uses, Interpretations, and Images*. New Brunswick: Transaction Books [First published 1987 in French as Rumeurs: Le Plus Vieux Média du Monde. Paris: Editions du Seuil]

Lewandowsky, S., Ecker, U. K., & Cook, J. (2017) Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of applied research in memory and cognition*, 6(4), 353-369.

Middleton, S.E., Lavorgna, A. McAlister, R. (2020) STAIDCC20: 1st International Workshop on Socio-technical AI Systems for Defence, Cybercrime and Cybersecurity. In: 1*2th ACM Conference on Web Science (WebSci '20 Companion)*.

Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D., & Nielsen, R. K. (2017) *Reuters Institute digital news report 2017*. Available at: SSRN 3026082.

NSPCC Learning (2021) *Using social media safely with children and young people*. Available from: https://learning.nspcc.org.uk/safeguarding-child-protection/social-media-and-online-safety.

ProTechThem (2021) *UKRI ESRC funded project exploring motivation for sharenting (parents sharing online information about minors) and automated detection of risk behaviours online*. Available from: https://www.protechthem.org/.

SafeSpacesNLP (2021) *UKRI TAS Hub funded project exploring behaviour classification natural language processing (NLP) in a socio-technical AI setting for online harmful behaviours for children and young people*. Avaibale from: https://www.tas.ac.uk/safespacesnlp/.

Shibutani, T. (1966) Improvised News: a Sociological Study of Rumours. Indianapolis:
Thomas, K. et al. (2021) SoK: Hate, Harassment, and the Changing Landscape of Online Abuse, *IEEE Symposium on Security and Privacy (SP)*.

Webb, H. and Jirotka, M. (2017) Nuance, Societal Dynamics and Responsibility in Addressing Misinformation in the post truth era: Commentary on Lewandowsky, Ecker and Cook. *Journal of Applied Research in Memory and Cognition,* 6(4), Dec 2017, pp. 414-417. doi.org/10.1016/j.jarmac.2017.10.001

Webb, H., Burnap, P., Procter, R., Rana, O., Williams, M., Stahl, B., Housley, W., Edwards, A., and Jirotka, M. (2016) Digital Wildfires: Propagation, Verification, Regulation and Responsible Innovation. *ACM Transactions on Information Systems, Special issue: Trust and Veracity of Information in Social Media*, 34(3), Article 15. doi.org/10.1145/2893478.

Whon, DY et al. (2017) How to Handle Online Risks? Discussing Content Curation and Moderation in Social Media. *CHI EA '17 Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. Pages 1271-1276

World Economic Forum (2013) *Digital Wildfires in a hyperconnected world. Global Risks Report. World Economic Forum*. Available from: http://reports.weforum.org/global-risks-2013/risk-case-1/digital-wildfires-in-a-hyperconnected-world/.

20 September 2021