Written evidence submitted by Rachel Coldicutt, OBE (OSB0153)

I am an independent expert on the social impact of new and emerging technologies and director of research consultancy Careful Industries.

Until 2019, I was founding CEO of responsible technology think tank Doteveryone, where I led influential research into how technology is changing society and developed practical tools for responsible innovation. Prior to that I spent almost 20 years working at the cutting-edge of new technologies for a range of media and technology companies, including a stint where I oversaw an online community of over 100,000 children and teenagers, and established many safety and wellbeing protocols for young people online. I am an advisor and board member for a number of organisations, and a non-executive member of the Ofcom Content Board, although it should be noted my response is independent and does not reflect the views of Ofcom in any way.

Doteveryone closed in 2020, and I am submitting evidence in part as a continuation of the regulatory research I oversaw in the four year period prior to that.

Will the proposed legislation effectively deliver the policy aim of making the UK the safest place to be online?

1. The majority of the measures set out in the Online Safety Bill focus on content and how to find it. However, content is only one part of the online experience and is not the only source of harm. If the UK is to be "the safest place to be online", the Bill should set out powers to scrutinise other components of apps and Web services, including their business models, Key Performance Indicators (KPIs), recommendation algorithms, and design patterns. To be effective, an online duty of care must begin as far upstream as possible in the product strategy and embrace a "safety by design" ethos; without doing so, the current precautionary approach to content publication risks being both onerous and instrumental.

2. At present, the legislation focuses mostly on content publication and discovery. As such the legislation risks (a) embedding a broadcast perspective in online regulation and (b) overweighting interventions relatively far downstream; the shelf life of these specific, content-centric, measures may be reduced by a number of factors, including the introduction of new business models and content discovery methods by platforms and online publishers, and the discovery of new social and individual harms.

3. It is also possible that the typologies of content and of harm enumerated in the Bill will not capture either emergent harms or harms with broad societal, rather than individual, consequences. This is in part because the sociotechnical impacts of technologies can unfold quickly, in surprising ways that shift cultural norms. Sometimes this is because a platform such as YouTube or TikTok offers scale and size to behaviours that have previously been hidden: ASMR videos, watching someone fix their washing machine, singing duets with strangers, and cheering as other people

play games online have all seemed remarkable within the last decade, but they are now everyday activities. The shifts that will occur in the coming decade are not easy to predict, and it is likely that significant harms will emerge that evade the current rubric.

4. As such, having a rights focus for the duty of care, modelled on the United Nations Guiding Principles on Business and Human Rights, rather a focus on upholding content standards will be more durable and easier to test against.

5. At the time of writing, new harms continue to come into public view on a daily basis – stories from the last week include a no-code deepfake pornography tool[1] and an algorithm that inordinately increases levels of in "Misinformation, toxicity, and violent content" on Facebook.[2] While these may seem wearily predictable, neither would necessarily fall in scope for the reporting processes recommended by the Bill.

6. While enumerating and recognising the variety of harms provoked by a small set of business models is important for the realisation of a healthy digital society, it is potentially also a problem-making exercise in the context of the legislation, as it detracts from recognising and effectively remedying the systemic causes. It is important that this legislation is actionable, sustainable, and creates levers for meaningful scrutiny; this means the focus must be specific but not overly granular and, vitally, not administratively gameable by sophisticated and labyrinthine platform companies.

7. This submission reminds the Committee of the importance of mandating greater transparency around the business models and KPIs and engagement metrics that drive online experiences, as well as the importance of building capacity for public redress, particularly for harms that emerge over time.

8. Possible upstream mitigations to reduce the level of harm generated by online business include mandatory transparency reporting around engagement models and Key Performance Indicators, so that platforms report to the regulator the target user behaviour they are working to achieve, and express the likelihood of this leading to harmful outcomes. A risk modelling tool such as Doteveryone's Consequence Scanning (now in the care of the Open Data Institute)[3] could be used to surface these outcomes at an early stage and mitigations put in place before products are released, rather than afterwards, when harm has already surfaced at notable scale. This puts ex-ante thinking at the heart of business and product strategy for the platforms, and would be a powerful lever for both greater transparency and healthier online experiences.

[1] Karen Hao, "A horrifying new AI app swaps women into porn videos with a click", *MIT Tech Review*, 13 September 2021

[2] Keach Hagey and Jeff Horwitz, "Facebook Tried to Make Its Platform a Healthier Place. It Got Angrier Instead", *Wall Street Journal*, 15 September 2021,

[3] Brown S. (2019) *Consequence Scanning Manual Version 1*. London: Doteveryone. https://doteveryone.org.uk/project/consequence-scanning/

9. It is also worth noting that not all "online harms" emerge fully fledged into the world, some emerge slowly through the interaction with the real world. This means that the risk and impact assessments as set out in the bill may have limited utility, particularly for some of the more complex and/or unimaginable uses of technology.

10. For instance, in 2014 bloggers Shafiqah Hudson and I'Nasah Crockett started to notice alt-right trolls posing as young Black women with extreme opinons on Twitter. This was later understood to be part of a campaign to grow political division that originated on 4Chan, and may have been related to a Russian propaganda campaign. Hudson and Crockett were able to recognise this behaviour as "sock puppeting", and began a social campaign, #YourSlipIsShowing, to bring it to public notice. The complex nexus of harms bundled in this campaign was wide-ranging, and included broad social harm to a protected class (specifically, damaging the credibility of young Black women), sowing social division, and potentially interfering in an election campaign. The social and linguistic clues that Hudson, Crockett and their peers detected over a period of months was subtle and likely beyond the scope of an algorithm to detect; revealing the campaign relied on the actions of an engaged community, working together. While not all of these harms would be within the current scope of the Online Safety Bill, this campaign has been recognised as a bellwether for the subsequent rise of alt-right propaganda campaigns in the wake of the 2016 US election and the related rise of polarising online content.[4]

11. Less dramatic examples can be found in the work of UK journalist Amelia Tait, who has documented many trends that have emerged through the ways content creators adapt their content and behaviour to serve both the needs of audiences and algorithms. Her articles on topics such as the manipulation of child influencers[5] and restricted and extreme eating show how platforms facilitate the creation of content that may not be considered harmful in isolation, but which – in aggregate, and once they become trends – have the potential to encourage harmful behaviour in viewers and other creators and would-be creators.

12. Flagging and observing these sociotechnical shifts may not always be the same as making a complaint about a piece of inappropriate content; instead, it speaks to a broader duty of care to monitor and understand the social impact of technology, spot the incentives that give rise to harmful behaviour, and improve standards and design patterns accordingly. One mechanism to achieve this may be through a shared regulatory front door[6] (perhaps for the Digital Regulation Cooperation Forum), where the public can raise concerns.

[4] Rachelle Hamilton, "The Black Feminists Who Saw the Alt-Right Threat Coming", *The Slate*, 23 April 2019

[5] Amelia Tait, "Are we failing to protect the child stars of YouTube?", *The New Statesman*, 10 November 2017

[6] Miller C, Ohrvik-Stott J, Coldicutt R. (2018) Regulating for Responsible Technology: Capacity, Evidence and Redress: a new system for a fairer future. London: Doteveryone. https://doteveryone.org.uk/project/regulating-for-responsible-technology/

13. On the point of child safety, it should also be noted that the possibility of absolute online safety for children always risks being compromised by both their proximity to adults and the existence of targeted ads and recommendation algorithms. Shared family devices and log-ins and IP level ad targeting all make it possible that children will be exposed to recommendations that relate to other people's activities. Something as everyday as watching YouTube on a parents' phone while they talk to a neighbour or finish an important task could easily result in seeing ads or recommendations for content that are not age appropriate.

14. Please note that I touched on the potentially harmful consequences of relying on algorithms to fulfil safety duties in recent written evidence to the House of Lords Communications and Digital Committee inquiry into Freedom of Expression Online. Those points remain substantive in the broader context of the Online Safety Bill.[7]

28 September 2021

[7] Rachel Coldicutt OBE—written evidence (FEO0122) House of Lords Communications and Digital Committee inquiry into Freedom of Expression Online. See points 7.7-7.15.