Rachel Coldicutt OBE—written evidence (FEO0122)

House of Lords Communications and Digital Committee inquiry into Freedom of Expression Online

2. How should good digital citizenship be promoted? How can education help?

2.1 This evidence submission makes the case that good digital citizenship requires a wider choice of platforms, supported by new models of ownership, and increased scrutiny of digital platforms' business strategies and internal success metrics. Educating users to be better digital citizens is one part of the puzzle, but for education to be effective, digital platforms must be held accountable for creating better civic spaces, and digital citizens must be entitled to exercise due rights and benefit from protections.

2.2 The concept of "good digital citizenship" is not easily defined, and any consideration of it must take into account the differences between life online and life offline.

2.3 Firstly, where is a digital citizen a citizen *of*? The history and ownership structures of "cyberspace" do not make it easy for individuals to translate the norms that may be expected offline into the online world. Digital governance is a complex topic and, while there is not space to give a complete overview here, it is worth noting that true "digital sovereignty" – the ability of a state or group of states to "act independently in the digital world"[1] – is not baked into the products and services that most people use day-to-day.

2.4 It is now generally accepted that a single global Internet has been replaced by the "Splinternet", a set of discrete socio-political technical territories.[2] In 2018, O'Hara and Hall identified four internets – and the cultural values of those four dominant territories define the global digital experience.

2.5 Libertarian Silicon Valley attitudes and behaviours have had outsized influence on the behaviour of UK digital citizens. The Centre for Data Ethics' recent polling reports that 66% of people in the UK use Facebook and YouTube, while 65% use WhatsApp. These platforms were not intentionally created to replace public or civic spaces around the world; in fact, their failure to adapt to and consider global social and cultural differences as they grew is an Internet-era example of colonisation, and they have continued to transmit the cultural values of their place of origin.

---

1 MADIEGA Tambiama, 'Digital Sovereignty for Europe', n.d., 12.

2 Kieran O'Hara and Wendy Hall, 'Four Internets: The Geopolitics of Digital Governance', Centre for International Governance Innovation, 7 December 2018, https://www.cigionline.org/publications/four-internets-geopolitics-digital-governance/.

2.6 The broader social impact of these technologies certainly took a back seat in early thinking about governance. For instance, the W3C still defines the World Wide Web as an "information space"[3] while John Perry-Barlow's 1996 speech "A Declaration of the Independence of Cyberspace" asserted that Cyberspace has "a natural independen[ce]" from government.[4]

2.7 While some local adaptations have been made to content and moderation policies (perhaps, most notably, with respect to Germany's NetzDG Act), the design of the platforms and underlying business models have remained much more globally consistent, with the result that we are all incentivised to be engaged *users* – helping product teams to achieve their Objective and Key Results (OKRs) – *before* we are given the tools to be good citizens. [5]

2.8 As such, even the most diligent and media-literate digital citizen must recognise and overcome the fact that the most popular digital platforms have not been created or optimised for a UK conception of digital citizenship. While education is undoubtedly necessary, it is not enough on its own to empower the majority of people to uphold standards and ways of being that run counter to the success metrics of the platforms they are using. The design, ownership, and strategic goals of the platforms being used have a bearing too.

2.9 Secondly, what is digital civic space? And is digital public space even possible without public ownership?

2.10 Civicus, the global civil society alliance, defines civic space as, "the bedrock of any open and democratic society. When civic space is open, citizens and civil society organisations can organise, participate, and communicate without hindrance. In doing so, they are able to claim their rights and influence the political and social structures around them."[6] This notion does not directly translate into the digital realm: while the platforms created by Facebook, Google, Snap, Byte Dance and Microsoft are often free to use at the point of access, they are not necessarily "open"; instead they are corporately owned spaces that require people to create accounts and consent to their behaviour being tracked, analysed and – ultimately – monetised.

2.11 Of course, social networks are not the only precincts in which public discourse can take place: video calling software, multi-player computer games, messaging services including Signal and Telegram, and collaboration tools such as Slack and Microsoft Teams all offer

---

3 'Help and FAQ - W3C', accessed 24 May 2021, https://www.w3.org/Help/#activity.

4 John Perry Barlow, 'A Declaration of the Independence of Cyberspace', Electronic Frontier Foundation, February 1996, https://www.eff.org/cyberspace-independence.

5 Karen Hao, 'He Got Facebook Hooked on AI. Now He Can't Fix Its Misinformation Addiction | MIT Technology Review', 11 March 2021, https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/.

6 'What Is Civic Space? - CIVICUS - Tracking Conditions for Citizen Action', accessed 24 May 2021, https://monitor.civicus.org/whatiscivicspace/.

opportunities for people to communicate in public, in private, and in the uncertain digital spaces in between the two.

2.12 Digital platforms' blurring of the boundaries between the public and private is another impediment to individuals aiming to be "good digital citizens". As noted in a recent study of on and offline behaviours in New York City during the pandemic, "Online spaces such as Facebook Groups or Zoom sessions often confounded boundaries of what people perceive as 'public' versus 'private.' People often viewed their personal social media accounts as private but participated in online communities they recognized as public or semi-public… Platform affordances contributed to—but did not determine—expectations of privacy and publicness".[7]

2.13 This report, "Terra Incognita NYC", defines digital public space as the result of civic action, and posits that it is possible to treat a commercially owned entity as public by simply using it as such. The paper defines publicness as a "sociopolitical practice", rather a condition of ownership, but this creates a vulnerability: the idea that a technology can simply be repurposed as public space with no democratic protections depends upon tacit consent between the technology providers and the technology users. In reality, terms of use, data collection, and feature sets can be changed at will by the provider, and the digital public has no immediate right to redress if the features that allow their "digital public space" to exist are changed or, simply, go away. The acquisition and subsequent change of terms of service for platforms including Instagram, Tumblr and Flickr have shown that privately owned platforms used as digital public infrastructure have few protections, and making good digital citizenship contingent on voluntary cooperation from private businesses is neither democratic nor sustainable.

2.14 Meanwhile, in the physical environment and in the media, it is acknowledged that a healthy mix of ownership and access models are required to support a healthy democracy and good outcomes for everyone.

2.15 Beyond work, home and school, "real life" is populated with a mix of spaces "where people come to congregate and linger regardless of what [or whether] they've purchased".[8] As well as "cafés, diners, barbershops and bookstores" the social infrastructure of a place includes, "public institutions, such as libraries, schools, playgrounds, parks, athletic fields, and swimming pools … sidewalks, courtyards, communities gardens … community organisations, including churches and civic associations". [9] Not all of life is expected, for instance, to take place in the mall, and not every action is intended to be monetised.

2.16 In the media landscape, the need for plurality has given rise to public service broadcasting (PSB). "PSB is an intervention, designed by

---

7 Mona Sloane, 'Terra Incognita NYC: A Study of NYC's Digital Public Space in the Pandemic' (New Public, 2021).

8 Eric Klinenberg, *Palaces for the People: How to Build a More Equal and United Society* (Bodley Head, 2018).

9 Ibid.

Parliament, to ensure that UK audiences can enjoy a wide range of high-quality programmes that meets people's needs as citizens and their interests as individuals… The PSB channels are all for public benefit, so are made available to all and are free at the point of delivery, without subscription or contract."[10] The purposes of PSB are set out in the Communications Act 2003.

2.17 But perhaps one of the most confounding aspects of the digital public space is that it does not map cleanly onto existing ideas about either the physical world or the media. While the regulatory focus is often on *content*, the focus of the platforms is *on influencing user behaviour*. This mismatch means that regulatory efforts can have, at best, a superficial impact on the way platforms operate; for instance, focusing on changing content and moderation policies will place restrictions on what people can say and do online, without removing the incentives to create of harmful and abusive content. While such measures might treat the symptom, they do not cure the underlying disease.

2.18 For example, Karen Hao's in-depth report into Facebook's use of artificial intelligence makes it clear that the company's business model is not optimised to support good digital citizenship: "A former Facebook AI researcher who joined in 2018 says he and his team conducted 'study after study' confirming the same basic idea: models that maximize engagement increase polarization. They could easily track how strongly users agreed or disagreed on different issues, what content they liked to engage with, and how their stances changed as a result. Regardless of the issue, the models learned to feed users increasingly extreme viewpoints. 'Over time they *measurably* become more polarized,' he says."[11]

2.19 This mismatch in ownership structures and democratic accountability in the digital realm means that good digital citizenship is expected to take place without the protections accorded in other aspects of life. Education is insufficient to overcome that, and interventions are needed to change the behaviour of the platforms, not just the users.

2.20 As Jillian C. York says in *Silicon Values*: "We've now firmly reached an era where it can no longer be said, if it ever could, that the internet is a space where "anyone, anywhere" may express their beliefs, but one in which groups already marginalised by society are further victimized by unaccountable platforms, and the already powerful are free to spread misinformation or hate with impunity."[12]

2.21 To create incentives for digital citizenship, citizens must be able to choose between publicly and privately owned digital platforms, and the regulation of existing platforms must encompass changes in product strategy and success metrics (or Objectives and Key Results, OKRs for short).

---

10 Ofcom, 'Small Screen: Big Debate – a Five-Year Review of Public Service Broadcasting (2014-18)', 27 February 2020, 67.

11 Hao, 'He Got Facebook Hooked on AI. Now He Can't Fix Its Misinformation Addiction | MIT Technology Review'.

12 Jillian C. York, *Silicon Values: The Future of Free Speech Under Surveillance Capitalism* (Verso, 2021).

2.22 US academic Ethan Zuckerman is an advocate for taxpayer-owned digital public infrastructure. "To put it simply, we need to imagine and build better digital public spaces that address the failures of our current infrastructures and actively work to create healthy and engaged civic discourse."[13] If digital spaces are an essential component of democracy, they must be treated as such; while good digital citizenship is an essential C21st competency, it is not achievable if the citizens have no rights over the digital spaces in which their citizenship takes place.

7. How can technology be used to help protect freedom of expression?

7.1 This submission will examine the use of technology to detect, monitor and otherwise moderate what the Online Safety Bill terms "legal but harmful" content.

7.2 One possible role technology can play in protecting freedom of expression is through algorithmic moderation to remove or minimise abusive behaviour and hate speech. This would, for instance, have the effect of liberating people with protected characteristics to participate fully in the online world and remove some barriers in for their participation high-profile careers.

7.3 The real-world effects of social-media abuse are becoming a limiting factor for offline freedom of expression and full participation in public life. The "No Excuse for Abuse" report outlines how online abuse is disproportionately targeted at people because of their profession or identity.[14]

7.4 Women MPs on Twitter and other social media are subject to high levels of abuse,[15] including threats of rape and murder; this has been a contributing factor in significant numbers of women stepping back from political life. Professor Anne Phillips has speculated that the wider "toxic atmosphere" of politics (including social media) will lead to women serving shorter terms as MPs than their male counterparts.[16]

7.5 Research by Amnesty International shows that Black women in the UK and US are 84% more likely to be mentioned in abusive or problematic tweets than white women, and 7.1% of all tweets sent to women examined in the study were problematic or abusive. "This amounts to 1.1 million tweets mentioning 778 women across the year, or one every 30

---

13 Ethan Zuckerman, 'What Is Digital Public Infrastructure?', *Center for Journalism and Liberty*, 17 November 2020, 20.

14 'No Excuse for Abuse', *PEN America* (blog), 31 March 2021, https://pen.org/report/no-excuse-for-abuse/.

15 Rosalynd Southern and Emily Harmer, 'Twitter, Incivility and "Everyday" Gendered Othering: An Analysis of Tweets Sent to UK Members of Parliament', Social Science Computer Review 39, no. 2 (1 April 2021): 259–75, https://doi.org/10.1177/0894439319865519.

16 LSE Government, Is a Toxic Atmosphere Driving Women Away from Politics?, 2019, https://www.youtube.com/watch?v=1RejxFHq4kY.

seconds. Women of colour, (black, Asian, Latinx and mixed-race women) were 34% more likely to be mentioned in abusive or problematic tweets than white women. Online abuse targets women from across the political spectrum - politicians and journalists faced similar levels of online abuse and we observed both liberals and conservatives alike, as well as left and right leaning media organizations, were targeted."[17]

7.6 The recent temporary boycott of social media by the FA, Premier League, EFL, FA Women's Super League, FA Women's Championship, PFA, LMA, PGMOL, Kick It Out, Women in Football and the FSA to protest racism and online discrimination further highlights the extent to which online hate speech has real-world repercussions.[18]

7.7 However, algorithms and large data sets are more likely to be biased against people with protected characteristics – the very people these moderation systems should be designed to protect. While the advantages of automating moderation processes are significant (and include reducing harm to content moderators), the challenges of executing algorithmic moderation in fair, just and effective ways require ongoing research, diligence and the creation of safeguarding standards and repeal processes. Complex and sensitive issues such as sexual and racial discrimination may not be easily detected by algorithmic processes, and the creation of large models to recognise and anticipate hateful content will require ongoing human management, governance, and oversight, and will pose cybersecurity and privacy risks.

7.8 Automated content moderation solutions have been used since the early days of Web forums. While the immediate context of an image or a sentence is important, so is the broader social and cultural context. This is often referred as "The Scunthorpe Problem", after residents of Scunthorpe were prevented from creating AOL accounts in 1996, as their addresses triggered the platform's swear filter.[19] While simple language filtering of this kind has improved significantly over the last 25 years, context remains everything: Facebook's longstanding "nipple ban" attracted significant criticism from breast cancer clinicians and breastfeeding specialists for assuming that all representations of nipples were pornographic. This policy was not updated until 2021, after being escalated to the Facebook Oversight Board.[20]

7.9 Prof. Safiya Noble's *Algorithms of Oppression* offers myriad examples of how "racism and sexism are part of the architecture and language of technology".[21] Algorithmic bias and the structural problems of using

---

17 'Why Twitter Is a Toxic Place for Women', accessed 24 May 2021, https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-1/.

18 'English Football Announces Social Media Boycott', accessed 24 May 2021, http://www.premierleague.com/news/2116111.

19 'Scunthorpe Problem', in *Wikipedia*, 21 May 2021, https://en.wikipedia.org/w/index.php?title=Scunthorpe_problem&oldid=1024267696.

20 Facebook, 'Oversight Board First Decision', 2021, https://about.fb.com/wp-content/uploads/2021/02/OB_First-Decision_Detailed_.pdf.

21 Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York University Press, 2018).

historic data to make decisions about the present or the future mean that there is a high propensity for data sets to replicate structural inequalities - including sexism, racism, ableism - in both the data they contain and the data they omit.

7.10 These problems do not go away with scale. In fact, they get worse. As Vinay Prabhu and Abeba Birhane state in their paper, "Large Datasets: A Pyrrhic Win for Computer Vision?", "Systems of classification, which operate within a power asymmetrical social hierarchy, necessarily embed and amplify historical and cultural prejudices, injustices, and biases… AI systems trained on such data amplify and normalize these stereotypes, inflicting unprecedented harm on those that are already on the margins of society."[22] *Algorithms of Oppression* gives examples of the "pornification of Black women", while the work of the Algorithmic Justice League has proven that white men are more likely to be correctly identified by facial recognition algorithms than Black women.[23]

7.11 Problems of bias are not restricted to images. The now-famous paper, "On the Dangers of Stochastic Parrots"[24] sets out the ways in which it can be difficult to scrutinise the contents of large training data sets, as well as being almost impossible to keep them up-to-date: "Given the compute costs of training large-scale models, it likely isn't feasible for even large corporations to fully retrain them to keep up with … language change."

7.12 Bender et al also show that men and women are likely to make different assessment of sexual harassment online – as such, could an algorithmic trained on data created or assessed by men be up to the job of detecting sexually harassing or abusive language? Training data for language models tends to come from aggregating existing large bodies text. Language change is not the only problem: the source of the training data is an issue too. Wikipedia is a popular source of training data for large-scale text models, but 85% of its contributors are men, so it is unlikely that the outcomes would be either neutral or favour women's interpretation.

7.13 While "Stochastic Parrots" largely considers text generation, rather than moderation, it does note that "components like toxicity classifiers would need culturally appropriate training data", and it seems very probable that the job of creating and stewarding exhaustive examples of offensive material for algorithms to learn from may well create problems than it might solve.

7.14 Meanwhile, GPT-2 – one of the largest and most diverse language models – has been shown to predict more stereotypical jobs for women than for

---

22 Vinay Uday Prabhu and Abeba Birhane, 'Large Image Datasets: A Pyrrhic Win for Computer Vision?', *ArXiv:2006.16923 [Cs, Stat]*, 23 July 2020, http://arxiv.org/abs/2006.16923.

23 Gender Shades study by Joy Buolamwini and Timnit Gebru. http://gendershades.org

24 Emily M. Bender et al., 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?;, in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21 (New York, NY, USA: Association for Computing Machinery, 2021), 610–23, https://doi.org/10.1145/3442188.3445922.

men, ultimately reflecting "the societal skew of gender and ethnicity in the US, and in some cases, pulls the distribution towards gender parity".[25] This shows how difficult it is to use machine-learning models to deal with cultural sensitivities, and raises questions as to whether or not universal models are possible.

7.15 As Gorwa et al note, "Algorithmic moderation has already introduced a level of obscurity and complexity into the inner workings of content decisions made around issues of economic or political importance, such as copyright and terrorism… A perfectly 'accurate' algorithmic moderation system would re-obscure not only the complex moderation bureaucracies that keep platforms functioning, but also threaten to depoliticise the fundamentally political nature of speech rules being executed by potentially unjust software at scale."[26] As such, deploying automated moderation to solve complex social problems and uphold the rights of people with protected characteristics would require a significant increase in transparency from digital platforms: allowing structural sexism and racism to be detected and remedied by "magic" algorithms would hand significant power to digital platforms, and may ultimately cause more problems than it will solve.

8. How do the design norms of platforms influence freedoms of expression? How can platforms create environments that reduce the propensity for online harms?

8.1 To reduce the propensity for online harms, platforms must adopt new metrics and business goals that prioritise good citizenship, trust and generosity above engagement. These goals must be published, and any new strategic plans should be run through an Equality Impact Assessments to show who is most likely to experience both harm and benefit from a particular technology or technological change.

8.2 Incremental design improvements, better privacy controls, and clearer terms and conditions will all lead to incremental improvements (see Caroline Sinders' *Trust Through Trickery* for more detailed design recommendations),[27] but these must be accompanied by more clarity on the goals of each platform. Every tweak made at every stage of product development is judged by whether it will help or hinder achieving the desired metrics; as such, doing anything other than changing these metrics is a cosmetic change.

May 2021

---

25 Hannah Kirk et al., 'How True Is GPT-2? An Empirical Analysis of Intersectional Occupational Biases', *ArXiv:2102.04130 [Cs]*, 8 February 2021, http://arxiv.org/abs/2102.04130.

26 Robert Gorwa, Reuben Binns, and Christian Katzenbach, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance', *Big Data and Society* 1–15, no. January-June (n.d.).

27 Caroline Sinders, Vandinika Shukla, and Elyse Voegeli, 'Trust through Trickery', *Commonplace*, 6 January 2021, https://doi.org/10.21428/6ffd8432.af33f9c9.