Written evidence submitted by LGBT Foundation (OSB0191)

We at LGBT Foundation are submitting the response to the enquiry launched by The Joint Committee on the Draft Online Safety Bill. LGBT Foundation is a national charity that works to support, amplify, and empower the LGBTQ+ community. We directly serve over 40,000 people annually as well as providing online support and advice to over 600,000 individuals, more than any other organisation of our size in the sector. Our work relies heavily on the ability of individuals to contact us online and express their needs in an open and transparent space, including discussing sensitive information that may be (incorrectly) seen as controversial or harmful.

Committee Question 5: Is the "duty of care" approach in the draft bill effective?

No, the "duty of care" principle needs to be amended to explicitly focus on illegal content. The inclusion of 'harmful' content in the bill without an explicit definition will effectively outsource decision making on what is and isn't permitted to tech companies. This will introduce avenues for discrimination against LGBT users and/or creators of LGBT-related content, with little recourse for appeal in the absence of specified standards.

The threat of large fines will create a general commercial incentive to over-censor, which existing evidence indicates will also lead to the disproportionate over-censorship of LGBT content as compared with other content. For example:

- The popular video-sharing social networking platform TikTok has an extensive history of discriminatory censorship of LGBT content, including limiting the reach of LGBT content in certain languages and temporarily removing the 'intersex' tag without explanation, making the content non-searchable for users.[1] In these instances, lack of transparency about how these decisions were made – including clear standards and the extent to which moderation is automated – has been cited as a major issue contributing to discriminatory censorship.
- In July 2021, the professional networking site LinkedIn removed a woman's post about her 16-year-old son coming out as gay and wearing a dress to his prom, also blocking her user profile.[2] This was reportedly due to some users flagging the post as inappropriate.

In relation to LGBT users and LGBT-related content, the inclusion of 'harm to children' under the "duty of care" approach without an explicit definition of what this constitutes is particularly dangerous, given significant evidence of non-explicit LGBT-related content being discriminatorily mis-classified as 'mature' or 'adult' by large platforms such as Tumblr and YouTube.[3] This resulted in young people being unable to access content about LGBT rights, history, identity, and discrimination, including in one case advice videos produced by an LGBT youth charity, while browsing in Restricted Mode on YouTube; and some LGBT content

---

[1] Kait Sanchez, "TikTok says the repeat removal of the intersex hashtag was a mistake: The tag's disappearance caused frustration for activists," *The Verge*, June 4, 2021, https://www.theverge.com/2021/6/4/22519433/tiktok-intersex-ban-mistake-moderation-transparency

[2] Tom Williams, "Brave teen came out to classmates by coming out in a dress for his prom," *Metro*, July 28, 2021, https://metro.co.uk/2021/07/28/linkedin-removed-mums-proud-post-of-her-son-coming-out-in-prom-dress-15004631/

[3] Claire Southerton, Daniel Marshall, Peter Aggleton, Mary Lou Rasmussen & Rob Cover. 2021. "Restricted Modes: Social Media, Content Classification and LGBTQ Sexual Citizenship." *New Media & Society*, 23:5, 920-38. https://doi.org/10.1177/1461444820904362.

and users being permanently prohibited from Tumblr after its decision to ban 'adult content' in December 2018 as a result of changes to US law. On YouTube, the restriction and demonetisation of videos mis-classified as 'mature' has directly resulted in financial harm to LGBT users, who in some cases have lost their livelihoods and/or been forced to stop creating content for the platform due to a decrease in advertising revenue. Notably, in late 2019 several LGBT YouTubers filed a class action lawsuit suing the platform for "discrimination, deceptive business practices and unlawful restraint of speech."[4] Though YouTube maintains a public list of guidelines for content demonetisation, the list is both broad and vague, meaning it is very difficult for users to understand what these standards are in practice and how they should be enforced.[5]

These examples show that LGBT content has already been disproportionately censored online when the decision over what might be "harmful" is left in the hands of private tech companies. The "duty of care" would not be effective in safeguarding LGBT users from discriminatory over-censorship; instead, by allowing companies to arbitrarily determine 'harm' within the framework of the bill, it would exacerbate and codify this censorship into law.

Committee Question 6: Does the bill deliver the intention to focus on systems and processes rather than content, and is this an effective approach for moderating content? What role do you see for e.g. safety by design, algorithmic recommendations, minimum standards, default settings?

A systems approach is the right one. However, the system in its current form incentivises companies to over-censor in order to avoid massive fines. This is a grave concern, given the prevalence of censorship of LGBT content on these platforms and their importance to LGBT users.

Systems approaches such as those specifically taken by YouTube and Tumblr have been shown to systemically discriminate against LGBT content. The 'safety by design' elements employed by Tumblr, such as defaulting to restricted mode and eventually banning all restricted content from the site, increased the severity of discriminatory over-censorship of LGBT content and reduced LGBT users' power to challenge this discrimination. YouTube's use of strategies such as automatic demonetisation of 'mature' content, and algorithmic recommendations that limit the reach of this content, has had a similar effect. This has resulted not only in the financial impact on LGBT creators described above, but the destruction of countless online spaces in which ordinary LGBT users – without significant individual platforms and therefore ability to challenge or appeal – previously found community and support.

These systems approaches are important and can in theory be used to promote genuine online safety, but good intentions result in discriminatory outcomes unless the appropriate

---

[4] Jenny Kleeman, "SNL producer and film-maker are latest to accuse YouTube of anti-LGBT bias: The 12 complainants in the class action lawsuit say an algorithm that restricts content is an attempt to push them off the platform," *The Guardian*, November 22, 2019, https://www.theguardian.com/technology/2019/nov/22/youtube-lgbt-content-lawsuit-discrimination-algorithm

[5] Aja Romano, "A group of YouTubers is trying to prove the site systematically demonetizes queer content: They reverse-engineered YouTube's ad revenue bot to investigate whether it's penalizing queer content," *Vox*, October 10, 2019, https://www.vox.com/culture/2019/10/10/20893258/youtube-lgbtq-censorship-demonetization-nerd-city-algorithm-report

care and attention are paid to countering the evident tendency towards discrimination within the existing design and implementation of these systems. As it stands, the bill does not adequately focus on the design of safe systems and processes. It must include written safeguards based on UK equality laws, developed in consultation with representative organisations, to ensure that equality and diversity is built into the design of these systems. Without these safeguards, it will lead to intensified discrimination against marginalised communities, including LGBT communities.

Committee Question 14: 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?

No, the definitions in the draft Bill do not include an explicit definition and process for determining what constitutes "harmful" content, and therefore afford too much discretion to individual service providers to distinguish "legal but harmful" vs "legitimate" content. As we have seen from the impact of existing moderation systems developed by a range of large service providers, this will result in significant discriminatory over-censorship of legitimate LGBT content and penalisation of LGBT users. Establishing effective, safe moderation systems in practice requires from the Bill, at minimum, a clear definition and process for determining harm to children and adults that incorporates principles of anti-discrimination.

Committee Question 19: 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?

Algorithms currently wrongly censor perfectly legal and safe content because they are less able to consider language nuances or context, and most programmes are grounded in datasets that incorporate discriminatory assumptions. LGBT communities, along with many other marginalised communities, have already felt the impact of discriminatory algorithms in recent years. In the YouTube and Tumblr examples above, the discrimination has been blamed by both companies on the design of their algorithms for moderating and classifying content. Despite this acknowledgement of issues with AI tools, there has been little change, with users facing the consistent problem of a lack of transparency about how these systems are designed.[6]

Algorithmically-driven discrimination is an issue for search engines as well as social media platforms. This is well-established within the academic literature, with a study on racial discrimination within search engine algorithms noting that "algorithmic oppression is not just a glitch in the system but, rather, is fundamental to the operating system of the web".[7] A recent study found that from 3rd-7th February 2020, "LGBT" and related search items in Google News "consistently provided a prominent platform for evangelical Christian and far-right perspectives on LGBTQ issues" rather than supportive perspectives, or LGBT-focused content from LGBT people themselves.[8] Meanwhile, a 2019 study conducted by CHEQ, an advertising verification company, found that advertising companies' keyword blacklists

[6] Ibid.

[7] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York, USA: New York University Press. https://doi.org/10.18574/9781479833641

[8] April Anderson & Andy Lee Roth. 2020. "Queer erasure: Internet browsing can be biased against LGBTQ people, new exclusive research shows." *Index on Censorship*, 49:1, 75-7. https://doi.org/10.1177/0306422020917088.

inappropriately targeted up to 73% of neutral or positive content from LGBT outlets such as *The Advocate* and *PinkNews*. The Bill must ensure that it does not incentivise and expand this overzealous, discriminatory algorithmic censorship.

Committee Question 20: Are there any foreseeable problems that could arise if service providers increased their use of algorithms to fulfil their safety duties? How might the draft Bill address them?

As demonstrated by the examples given above, the increased use of algorithms within the framework of the Bill would have a profound negative impact on LGBT users. The vague wording of the "Duty of Care," combined with the threat of hefty fines, would exacerbate and codify into law the issues we have already witnessed with over-censorship. The Bill would necessitate a vast increase in the use of algorithms for platform-wide content moderation which, in turn, would likely lead to a mass deletion and restriction of LGBT content. Many LGBT people would be blocked and removed from online platforms, making them unable to use certain social platforms and services and isolating them from the wider LGBT community, as well as destroying whole online LGBT user communities that have formed on particular platforms.

The loss of this online space would be harmful for LGBT people for whom the internet is an important space. Prior to the prohibition of 'mature' content from Tumblr in December 2018 resulting in a large migration away from the site, many LGBT young people identified social media platforms such as Tumblr as a prominent space in the formation of their sexual identity.[9] UK-based research also shows that 90% of LGBT young people say they can be themselves online and 96% say the internet has helped them understand more about their sexual orientation and/or gender identity.[10] Losing these positive aspects of the internet for the community, as a place to form connections, access vital resources and share experiences, would be particularly damaging for young LGBT people and those who are already the most socially vulnerable.

However, addressing the problems arising from increased use of algorithms is more complex than simply introducing human moderation as well. While escalating 'complex cases' to human moderators has been suggested as a potential mitigation, similar issues with discrimination have plagued Facebook's existing hybrid AI-human moderation and review system. In 2018, we saw Facebook wrongfully block adverts containing LGBTQ+ content due to them falling under the category of 'political' content, even though it did not contain any advocacy or political views.[11] The company also has a broader history of inappropriately blocking LGBT advertisements[12], and of enforcing its 'hate speech policies' in such a way that the system instead negatively targets the marginalised users such policies

---

[9] Alexander Cho. 2018. "Default publicness: Queer youth of colour, social media, and being outed by the machine." *New Media & Society*, 20:9, 3183-200. https://doi.org/10.1177/1461444817744784

[10] Josh Bradlow, Fay Bartram, April Guasp & Vasanti Jadva. 2017. *School Report: The experiences of lesbian, gay, bi and trans young people in Britain's schools in 2017.* London: Stonewall. https://www.stonewall.org.uk/school-report-2017

[11] Eli Rosenberg, "Facebook blocked many gay-themed ads as part of its new advertising policy, angering LGBT groups," *The Washington Post*, October 3, 2018. https://www.washingtonpost.com/technology/2018/10/03/facebook-blocked-many-gay-themed-ads-part-its-new-advertising-policy-angering-lgbt-groups/

[12] Sera Golding-Young, "Facebook's Discrimination Against the LGBT Community: The company rejected an ad of a same-sex couple, but took no issue with a similar ad of a heterosexual couple," *ACLU*, September 24, 2020, https://www.aclu.org/news/lgbtq-rights/facebooks-discrimination-against-the-lgbt-community/

were intended to protect.[13] This is due to a combination of insufficient understanding of how marginalisation and discrimination operate on the part of company leadership; inadequate one-size-fits-all policies that nevertheless are inconsistently applied; lack of transparency and accountability regarding how the systems operate and how decisions are made, with insufficient mechanisms for individual users to meaningfully appeal decisions; and a consistent lack of adequate resources committed by service providers to addressing these issues. These issues must be addressed holistically to ensure service providers develop systems, involving the use of algorithms where needed, that genuinely protect users without discrimination.

In particular, the Bill should consider in greater detail the mechanisms for individual users to appeal against the censorship of their content, and how the Bill might create additional incentives for platforms to remove legitimate content swiftly. The LinkedIn incident described above is only one example of complaints and reporting mechanisms being abused in a homophobic or transphobic manner to target LGBT content, creating new means for harassment of LGBT individuals. While the Bill places the onus platforms to swiftly remove potentially "harmful" content under threat of significant fines, without greater consideration of the potential for discriminatory abuse of these systems users will be left even more vulnerable to such forms of harassment.

In summary, while the Bill is well-intentioned and seeks to mitigate potential issues, as it stands the safeguards on freedom of expression and routes for user appeal outlined within it are inadequate to the foreseeable scope, character, and complexity of the problem of discriminatory over-censorship of LGBT users. The consequences of potential 'under-censorship' for service providers would be costly, while the cost of discriminatory over-censorship would be low – resulting in a high risk that providers default to the removal of online content or suspension of accounts that are legitimate, and that content will be removed in accordance with biased or discriminatory concepts. Service providers' positions as 'control points' shaping how information is accessed and which voices are heard means that such discrimination would, conversely, significantly harm LGBT users and communities.[14] In addition to providing an explicit definition and process for determining "harm" and appropriate direction for building anti-discrimination principles into moderation systems, the Bill fundamentally must embed these within a regulatory framework that truly safeguards LGBT users from harm by incentivising service providers to understand discriminatory over-censorship as a serious problem, and invest resources in addressing it.

28 September 2021

---

[13] Dottie Lux & Lil Miss Hot Mess, "Facebook's Hate Speech Policies Censor Marginalized Users," *Wired*, August 14, 2017, https://www.wired.com/story/facebooks-hate-speech-policies-censor-marginalized-users/

[14] Laura DeNardis & Andrea M. Hackl. 2016. "Internet control points as LGBT rights mediation." *Information, Communication & Society*, 19:6, 753-70. https://doi.org/10.1080/1369118X.2016.1153123