Twitter—supplementary written evidence (FEO0121)

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

Thank you again for inviting Twitter to participate in this inquiry, and the opportunity to provide further written evidence on Birdwatch, Competition, Project Bluesky and Online Harms. Given the Committee's interest in the topic, we have also included our approach to pseudonymous accounts.

Birdwatch

People come to Twitter to stay informed, and they want credible information to help them do so. We apply labels and add context to Tweets,[1] but we don't want to limit efforts to circumstances where something breaks our rules or receives widespread public attention. We also want to broaden the range of voices that are part of tackling this problem, and we believe a community-driven approach can help. That's why we're piloting Birdwatch,[2] a new community-driven approach to help address misleading information on Twitter.

Birdwatch allows people to identify information in Tweets they believe is misleading and write notes that provide informative context. We believe this approach has the potential to respond quickly when misleading information spreads, adding context that people trust and find valuable. Eventually we aim to make notes visible directly on Tweets for the global Twitter audience, when there is consensus from a broad and diverse set of contributors.

In this first phase of the pilot, notes will only be visible on a separate Birdwatch site.[3] On this site, pilot participants can also rate the helpfulness of notes added by other contributors. These notes are being intentionally kept separate from Twitter for now, while we build Birdwatch and gain confidence that it produces context people find helpful and appropriate. Additionally, notes will not have an effect on the way people see Tweets or our system recommendations.

We have conducted more than 100 qualitative interviews with individuals across the political spectrum who use Twitter, and we received broad general support for Birdwatch. In particular, people valued notes being in the community's voice (rather than that of Twitter or a central authority) and appreciated that notes provided useful context to help them better understand and evaluate a Tweet (rather than focusing on labeling content as "true" or "false"). Our goal is to build Birdwatch in the open, and have it shaped by the Twitter community.

---

1 https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html

2 https://twitter.com/i/birdwatch

3 https://twitter.com/i/birdwatch

To that end, we're also taking significant steps to make Birdwatch transparent:

- All data contributed to Birdwatch will be publicly available and downloadable in TSV files[4]

- As we develop algorithms that power Birdwatch - such as reputation and consensus systems - we aim to publish that code publicly in the Birdwatch Guide.[5] The initial ranking system for Birdwatch is already available.[6]

We hope this will enable experts, researchers, and the public to analyse or audit Birdwatch, identifying opportunities or flaws that can help us more quickly build an effective community-driven solution.

Competition, Project Bluesky and Online Harms

We have welcomed the opportunity to engage in the Online Harms consultation process. Specifically, we believe regulatory frameworks that look at system-wide processes, as opposed to individual pieces of content, will be able to better reflect the challenges of scale that modern communications services involve - and we therefore welcome the government's stated commitment to this approach. Indeed, we are also pleased to be members of the new Digital Trust and Safety Partnership.[7] A first-of-its-kind partnership, we are committed to developing industry best practices, verified through internal and independent third-party assessments, to ensure consumer trust and safety when using digital services.

Similarly, we are also supportive of Ofcom's designation as the regulator for Online Harms. As we stated in our submission to the White Paper back in 2019, we believe Ofcom is the most appropriate and qualified body to be designated as the independent regulatory authority.

We do believe, however, that there are opportunities to provide further clarity - and better future-proof - the regulation. Competition, for example, is critical for our industry to thrive; we believe that the Open Internet is at risk of being less open as it becomes less competitive.

Globally, we are urging regulators to factor into their decisions a test of whether proposed measures further enhance the dominance of existing players; or set insurmountable compliance barriers and costs for smaller companies and new market entrants.

We believe a forward-thinking approach might also consider the adoption of technical standards that embed openness and interoperability. Interoperability protects both public choice and competition and is a policy tool that has significant

---

4 https://twitter.com/i/birdwatch/download-data
5 https://twitter.github.io/birdwatch/
6 https://twitter.github.io/birdwatch/about/ranking-notes/
7 https://dtspartnership.org/

potential.

To accelerate efforts here, Twitter is funding a small independent team of open source architects, engineers, and designers to develop an open and decentralised standard for public conversation on the Internet. We call this Project Bluesky.[8] Ultimately, the hope is that Twitter will be one of many clients of this standard and it can become the bedrock of a new, democratised way of building services that can connect us. This team will not only develop a decentralised standard for public conversation, but also build an open community around it - inclusive of companies and organisations, researchers, and civil society leaders - all of whom are thinking deeply about the consequences, positive and negative.

If successful, the adoption of this standard could allow us to access and contribute to a much larger corpus of public conversation in the long term. At Twitter, we would then be able focus our efforts on building open recommendation algorithms which promote healthy conversation and access to credible content.

In the meantime, we believe the regulatory debate needs to reflect how content moderation is now more than just leaving content up or taking it down. The fundamental question of how people find content - and how it is amplified - is more important than just where and how content exists. It can no longer be seen as a binary debate.

There are a number of areas where the present proposals for 'legal but harmful' content may cause confusion for Internet users. To offer just one example of the challenges of this approach, the proposed exemption for journalists in the Online Harms Full Response may have unintended consequences. The Full Response states: "*legislation will include robust protections for journalistic content shared on in-scope services.*" Journalism is the lifeblood of Twitter - we believe in it, we advocate for it, and we seek to protect it. What's more, we recognise[9] that sometimes it may be in the public interest to allow people to view Tweets that would otherwise be taken down, and have developed policies and processes accordingly. The challenge with translating this to regulation is the absence of a clear definition of what constitutes 'journalistic content.' Every day we see Tweets with screenshots of newspaper front pages, links to blogs, updates from journalists and firsthand accounts of developing events. Crucially, there are accounts we have suspended for Hateful Conduct and other violations of our rules who have described themselves as 'journalists.' If the Government wishes for us to treat this content differently to other people and posts on Twitter, then we would ask the Government to define it, through the accountability of the Parliamentary process. Without doing so, it risks confusion not just for news publishers and for services like ours, but for the people using them.

As well as the regulatory focus on content types and moderation systems, we believe it is essential that consumers are given more control over the algorithms

---

8 https://twitter.com/bluesky

9 https://help.twitter.com/en/rules-and-policies/public-interest

that shape their online experience as part of a forward-looking policy approach. Twitter's decision to include such a control on our home timeline highlights how users can be empowered without undermining the service. Our interdisciplinary Responsible Machine Learning working group[10] are in the early stages of exploring this further - algorithmic choice will allow people to have more input and control in shaping what they want Twitter to be for them.

Openness and transparency will remain core to Twitter's approach. Transparency is embodied in our open APIs, our information operations archive, and our disclosures in the Twitter Transparency Center.[11] We continue to encourage policymakers, including in the UK, to expand transparency requirements to enable greater accountability. We've gone further than our peers, offering a new academic platform to encourage cutting edge research using Twitter data, for instance, and a Covid-19 endpoint to empower public health research. Over 100 researchers and developer teams representing 30 different countries were granted access to the Covid-19 API stream endpoint. More than half of them focused on studying disinformation and misinformation around Covid-19; others examined public perceptions, sentiment, and the evolution of people's attitudes about the pandemic over time. This is one of the reasons you hear more about reports featuring Twitter as core to the research methodology - we enable and empower it.

Over the coming months, we plan to build on this further. We're conducting in-depth analysis and studies to assess the existence of potential harms in the algorithms we use. Here are some analyses you will have access to in the upcoming months:[12]

- A gender and racial bias analysis of our image cropping (saliency) algorithm (published May 2021)[13]
- A fairness assessment of our Home timeline recommendations across racial subgroups
- An analysis of content recommendations for different political ideologies across seven countries

Pseudonymity

At Twitter, we are guided by our values, and never more so than when it comes to fundamental issues like identity.

---

10 https://blog.twitter.com/en_us/topics/company/2021/introducing-responsible-machine-learning-initiative.html

11 https://transparency.twitter.com/

12 https://blog.twitter.com/en_us/topics/company/2021/introducing-responsible-machine-learning-initiative.html

13 https://blog.twitter.com/engineering/en_us/topics/insights/2021/sharing-learnings-about-our-image-cropping-algorithm.html

Pseudonymity is no shield against our policies - we employ algorithms which are constantly looking to find, challenge and remove accounts breaking any of our rules. This might include misleading others through the use of a fake account, such as through intentionally misleading profile information, location, and/or the use of stock or stolen profile photos.

When you sign up for an account, we ask you for your name, your date of birth and your phone number or email address, which you must verify in order to continue with account creation.

Indeed, the police can contact us through our online portal to make requests about an account anytime.

We believe, however, that everyone has the right to share their voice without requiring a government ID to do so. Our approach in this space has been developed in consultation with leading NGOs - while pseudonymity has been a vital tool for speaking out in oppressive regimes, it is no less critical in democratic societies.

Pseudonymity may be used to explore your identity, to find support as victims of crimes, or to highlight issues faced by vulnerable communities. Indeed, many of the first voices to speak out on societal wrongdoings, have done so behind some degree of pseudonymity - once they do, their experience can encourage others to do the same, knowing they don't have to put their name to their experience if they're not comfortable doing so. It was often pseudonymous accounts that brought the first images out of Wuhan and Xinjiang at the beginning of the pandemic.

Perhaps most fundamentally of all - some of the communities who may lack access to government IDs are exactly those who we strive to give a voice to on Twitter. Estimates have suggested there are 3.5 million people[14] in the UK who don't have access to official forms of photo ID.

It is also important to remember that the UK government actually did seek to mandate ID verification for pornography websites; a policy dropped two years after the law was passed, after repeated delays and criticisms it would not work.

It is challenging to envisage how such a requirement for social media services might be globally scalable. Mandating ID would most immediately have the effect of disenfranchising billions of people from countries who do not have robust ID frameworks. Alternatively, were the law targeted at UK users, as was the case with pornography websites, technoloogy such as virtual private network services - which make it seem like a computer based is located in a different country - would quickly enable users to bypass the law. Either way, any such requirement would also require technology companies to store and process vast amounts of sensitive personal data - another concern that was raised when ID verification for

14 https://commonslibrary.parliament.uk/voter-id-key-facts-and-figures/

pornography websites was being developed.

We welcome the opportunity to engage with the Committee on these key questions - please let me know if you have any further questions.

May 2021