Written Submission of Karen White
Twitter International Company
Oireachtas Committee on Justice and Equality
9 October 2019

I would like to thank the Committee for your invitation to Twitter to participate in today's session. My name is Karen White, I am Director of Public Policy for Twitter in Europe.

Firstly, I will outline some key elements of our efforts to combat online harassment, and then provide our observations on the Harassment, Harmful Communications and Related Offences Bill.

Twitter is an open, public service. Our singular mission is to serve the public conversation. We serve our global audience by focusing on the needs of the people who use our service, and we put them first in every step we take. We must be a trusted and healthy place that supports free and open democratic debate.

Twitter is committed to improving the collective health, openness, and civility of public conversation on our platform. Our success is built and measured by how we help encourage more healthy debate, conversations, and critical thinking. Conversely, abuse, malicious automation, and manipulation detracts from it. We are committing Twitter to hold ourselves publicly accountable towards progress.

Controlling the Twitter Experience

We provide people on Twitter with a range of tools so that they can control their experience on our platform and manage the types of content and accounts they see.

Protected Tweets. When individuals sign up for Twitter, their Tweets are public by default; anyone can view and interact with public Tweets. People who use Twitter can choose to protect their Tweets. Those individuals with protected Tweets will receive a request when new people want to follow them, which they can approve or deny. Only approved followers are able to view an individual's protected Tweets, direct a Tweet at the individual, or send them a Direct Message.

Unfollowing Accounts. If individuals no longer wish to see the Tweets of someone they are following, people can select "unfollow" and the Tweets will no longer be visible on their

home timeline. The Tweets remain available to other followers, and can be viewed on an as-need basis by visiting the profile.

Blocking. Individuals on Twitter can "block" other accounts. This tool controls the way in which an individual interacts with other accounts on Twitter. When an account is blocked, a whole range of consequences arise to protect the user. Some of these include the inability of the blocked account to contact or follow the individual. Other consequences include adding the individual's Twitter account to the list of blocked account list when the block function is used.

Mute. An additional tool that Twitter provides to individuals who use our platform is the "mute" feature that allows an individual to remove an account's Tweets from the individual's timeline without unfollowing or blocking that account. Muted accounts can continue to follow the individual and the individual can continue to follow muted accounts. Muting an account will not cause an individual to unfollow them and the muted account is able to send a Direct Message to the individual. Replies and mentions by the muted account will appear in Notifications tab of the individual, but Tweets from a muted account – posted before the account was muted – will be removed from the individual's home timeline. If the individual clicks or taps into a conversation, replies from muted accounts will be visible.

Safe Search. There are many ways to use search on Twitter. Individuals can find Tweets from themselves, friends, local businesses, and everyone from well-known entertainers to global political leaders. By searching for topic keywords or hashtags, individuals can follow ongoing conversations about breaking news or personal interests. We give people control over what they see in search results through safe search mode. These filters exclude potentially sensitive content from the search results, such as spam, adult content, and the accounts an individual has muted or blocked. Individual accounts may mark their own posts as sensitive as well. Safe search is enabled by default, and people have the option to turn safe search off, or back on, at any time. This tool is explained in greater detail below.

Notifications Filters. Twitter's notification timeline offers users a simple way to see how others on Twitter are interacting with the user. From the notifications timeline, users are able to see which of their Tweets have been liked, plus the latest Retweets (of their Tweets), Tweets directed to them (replies and mentions) and the users' new followers.

We give individuals on Twitter additional controls over the content that appears in the notifications timeline, since notifications may contain content an individual on Twitter has not directly chosen to engage with (such as mentions or replies from someone the individual does not follow). By default, we filter notifications for quality, and exclude notifications about duplicate or potentially spam Tweets. We also give individuals on the platform granular controls over specific types of accounts they might not want to receive notifications from, including new accounts, accounts the individual does not follow, and accounts without a confirmed phone or email address.

Online Abuse

Twitter strives to provide an environment where people can feel free to express themselves. If abusive behaviour happens, Twitter wants to ensure that it is easy for people to report it to us. In order to ensure that people feel safe expressing diverse opinions and beliefs, Twitter prohibits behaviour that crosses the line into abuse, including behaviour that harasses, intimidates, or uses fear to silence another's voice.

We recommend that individuals on Twitter unfollow and end any communication with that account or block an account to prevent that person from following the individual, seeing the individual's profile image on their profile page, or in their timeline. Abusive accounts often lose interest once they realise that an account will not respond. Twitter recommends that if the account in question is a friend, try addressing the issue offline.

Anyone can report abusive behaviour directly from a Tweet, profile, or Direct Message. An individual navigates to the offending Tweet, account, or message and selects an icon that reports that it is abusive or harmful. Other options are available, for example posting private information or a violent threat. Multiple Tweets can be included in the same report, helping us gain better context while investigating the issues to resolve them faster. For some types of report Twitter also prompts the individual to provide more information concerning the issue that is being reported.

Twitter uses a combination of machine learning and human review to adjudicate abuse reports and whether they violate our rules. Context matters when evaluating abusive behaviour and determining appropriate enforcement actions. Factors we may take into consideration include, but are not limited to whether: The behaviour is targeted at an individual or group of people; the report has been filed by the target of the abuse or a bystander; and the behaviour is newsworthy and also in the legitimate public interest. Twitter subsequently provides follow-up notifications to the individual who reports the abuse. We also provide recommendations for

additional actions that individuals can take to improve their Twitter experience, for example using the block or mute feature.

Hateful Conduct Policies

We recognise that if people experience abuse on Twitter, it can jeopardise their ability to express themselves. Research has shown that some groups of people are disproportionately targeted with abuse online. This includes; women, people of colour, lesbian, gay, bisexual, transgender, queer, intersex, asexual individuals, marginalised and historically underrepresented communities. For those who identify with multiple underrepresented groups, abuse may be more common, more severe in nature and have a higher impact on those targeted.

We are committed to combating abuse motivated by hatred, prejudice or intolerance, particularly abuse that seeks to silence the voices of those who have been historically marginalised. For this reason, we prohibit behaviour which targets individuals with abuse based on a protected category.

An individual using the platform is not permitted to promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories.

We do not allow individuals to use hateful images or symbols in their profile image or profile header. Individuals using the platform are not allowed to use their username, display name, or profile bio to engage in abusive behaviour, such as targeted harassment or expressing hate towards a person, group, or protected category.

Under this policy, we take action against behaviour that targets individuals or an entire protected category with hateful conduct. Targeting can happen in a number of ways, for example, mentions, including a photo of an individual, referring to someone by their full name, etc.

When determining the penalty for violating this policy, we consider a number of factors including, but not limited to the severity of the violation and an individual's previous record of rule violations. For example, we may ask someone to remove the violating content and serve a period of time in read-only mode before they can Tweet again. Subsequent violations will lead to longer read-only periods and may eventually result in permanent account suspension. If an

account is engaging primarily in abusive behaviour, or is deemed to have shared a violent threat, we will permanently suspend the account upon initial review.

Twitter's Efforts to Combat Self-Harm and Suicide

Twitter informs individuals using our service:

> "If you or someone you know is at risk of self-harm or suicide, you should seek help as soon as possible by contacting agencies specialising in crisis intervention and suicide prevention. Also alert the team devoted to handling threats of self-harm or suicide if you encounter such threats on Twitter."

After we assess a report of self-harm or suicide, Twitter will contact the reported user and let the reported userknow that someone who cares about them identified that they might be at risk. We will provide the reported user with available online and hotline resources and encourage them to seek help.

Judging behaviour based on online posts alone is challenging, but there are potential warning signs or indicators for self-harm or suicide. We provide individuals using Twitter a list of questions to ask to help assess if another user is feeling suicidal:

- Does this person post content about depression or feelings of hopelessness?
- Is this person posting comments about death or feelings that death is the only option?
- Is this person posting comments about having attempted suicide in the past?
- Is this person describing or posting photos of self-harm or identifying him or herself as suicidal?
- Has his or her mood and the content of his or her posts changed recently?

We recommend that if the Twitter user knows the person involved, the individual should encourage him or her to seek professional help. We also inform individuals on Twitter if they are having thoughts of self-harm, suicide, or depression, please reach out to someone and request help. We provide them links to resources on depression, loneliness, substance abuse, illness, relationship problems, and economic problems.

Importantly, Twitter currently has in place product interventions that address potential self-harm. In response to certain keyword searches relating to self-harm, we direct individuals using Twitter's search function to online prevention resources. This feature connectes people with resources when they searched certain terms related to suicide and self-harm. This is available in Ireland, where we have partnered with the Samaritans and direct individuals to their

website and support services. Additionally, we have recently made updates to our reporting flow globally, to do our part in reducing the stigma around suicide. Now, reporting a Tweet that suggests someone intends to hurt themselves will no longer need to be reported as abusive content.

Twitter's Policies on Non-Consensual Nudity

Twitter does not allow individuals on the service to post or share intimate photos or videos of someone which were produced or distributed without their consent. We inform our users that sharing explicit sexual images or videos of someone online without their consent is a severe violation of their privacy and the Twitter Rules. Sometimes referred to as revenge porn, this content poses serious safety and security risks for people affected and can lead to physical, emotional, and financial hardship.

Under this policy, individuals on Twitter cannot post or share explicit images or videos that were taken, appear to have been taken or that were shared without the consent of the people involved. Examples of the types of content that violate this policy include, but are not limited to:

- hidden camera content featuring nudity, partial nudity, and/or sexual acts;
- creepshots or upskirts - images or videos taken of people's buttocks, up an individual's skirt/dress or other clothes that allows people to see the person's genitals, buttocks, or breasts;
- images or videos that superimpose or otherwise digitally manipulate an individual's face onto another person's nude body;
- images or videos that are taken in an intimate setting and not intended for public distribution; and
- offering a bounty or financial reward in exchange for intimate images or videos.

Results of Our Efforts

We have made meaningful progress in creating a healthier service. Since we announced our focus on improving the health of the conversation occurring on Twitter, we have seen:

- A 38 percent of the abusive content where Twitter takes action to enforce the removal from our platform is surfaced by our teams proactively for review, instead of relying on reports from people on Twitter.
- A 16 percent year-over-year decrease in reports from people complaining about interactions with other users allegedly abusing them on Twitter. Moreover, enforcement on reported content that was three times more effective.

- There was a 45 percent increase in the number of account suspensions for those users who attempted to create new accounts following their original suspension. Over the first quarter of 2019 this amounted to over 100,000 account suspensions for these "reoffenders".
- 60 percent faster response to appeals requests with our new in-app appeal process.
- 3 times more abusive accounts suspended within 24 hours after a report compared to the same time last year.
- 2.5 times more private information removed with a new, easier reporting process.

People who don't feel safe on Twitter shouldn't be burdened to report abuse to us. Previously, we only reviewed potentially abusive Tweets if they were reported to us. We know that is not acceptable, so earlier this year we made it a priority to take a proactive approach to abuse in addition to relying on people's reports.

Today, by using technology, 38 percent of abusive content that's enforced is surfaced proactively for human review instead of relying on reports from people using Twitter. This encompasses a number of policies, such as abusive behaviour, hateful conduct, encouraging self-harm, and threats, including those that may be violent.

The same technology we use to track spam, platform manipulation and other rule violations is helping us flag abusive Tweets to our team for review. With our focus on reviewing this type of content, we've also expanded our teams in key areas and geographies so we can stay ahead and work quickly to keep people safe. Reports give us valuable context and a strong signal that we should review content, but we've needed to do more and though still early on, this work is showing promise.

Legislation on Harassment, Harmful Communications and Related Offences

In order to ensure people can continue to express themselves freely and safely on Twitter, we know we must continue investing further in proactive technology, safety tools, as well as developing policies which keep pace with the changing contours of public conversation. With hundreds of millions of Tweets posted every single day, this is challenging but no more challenging than the task faced by this Committee in addressing the complex issue of harmful communications.

The Committee has asked for our recommendations on the proposed legislation. It is important that the legislation is as consistent as possible with existing legal frameworks to avoid uncertainties and discrepancies. This is particularly pertinent with reference to Articles 12 to 15 of the European Union e-Commerce Directive and Article 25 of Directive 2011/93/EU.

The effectiveness of any legislative solution relies on it being proportionate, technically feasible, and flexible, particularly given the diversity of companies within the digital ecosystem. In this context, the Committee will need to consider and assess how different legal and illegal harms manifest themselves across different platforms and varied jurisdictions.

We all share the objective of protecting our systems of due process and our commitment to freedom of expression. Preserving these tenets in regulatory proposals can be achieved by collectively ensuring there is clarity on the obligations of all stakeholders, thereby avoiding an outcome whereby companies could overreach and erroneously remove content that should otherwise be kept online. A clearly defined scope will assist Twitter and others.

We stand ready to work with the Committee as we continue to explore options to ensure that all people are protected from online harassment and harmful communications.