Twitter - supplementary written evidence (DAD0103)

Dear Lord Puttnam,

Follow up from Twitter: Hearing on 17 March 2020

Thank you again for inviting Twitter to provide evidence to the Select Committee on Democracy and Digital Technologies. We appreciated the opportunity to participate. During the hearing, I noted that I would follow up with the Committee on several areas.

First, we were asked about providing examples of Tweets that break our rules. For content moderators, our training material includes examples that illustrate policy violations, most of which are either hypothetical or reframed to remove private information. On our website, we provide a number of examples of Tweets that would violate our rules (for instance, information about our Hateful Conduct and Abuse policies). We also include such examples when we launch new policies or update existing ones - most recently, for instance, our blog post and proactive communications on our updated Dehumanization policy included a series of example Tweets.

Second, we were asked to provide further information regarding our content moderator training programme. Throughout the training phase, there are several certification exams that our content moderators must pass with a strict grading system. Part of this examination process checks for quality and accuracy in moderation, and helps us to identify any risk for bias and/or false positives. There is also regular on-job training provided and spot-checks to ensure a high standard of moderation at all times. The wellbeing of those who review content is a primary concern for our teams and our highest priority is to ensure our staff and partners are treated with compassion, care, and respect.

Third, you asked how many users had experimented with hidden replies. While we have not currently provided the overall percentage of all users in Japan or Canada who used the feature during the trial, we have provided the findings of the trial on our website here. We have also announced that we will be soon be launching a new hide replies endpoint in our API - this will enable external developers to build additional conversation management tools. We welcome opportunities to collaborate to promote this feature, and continue to promote it ourselves on an ongoing basis - see here for an example from this week.

Fourth, during the Committee hearing, I was asked if a specific Tweet would violate our rules. I have been unable to locate the Tweet referenced, but if you wish to send through a link to the Tweet we would be very happy to review if there have been any violations of our rules.

Finally, I said I would provide a timeline of the safety changes we have made at Twitter. Please see that document attached.

Thank you again for inviting Twitter to participate. Do get in touch if there are any further questions that we can assist with.

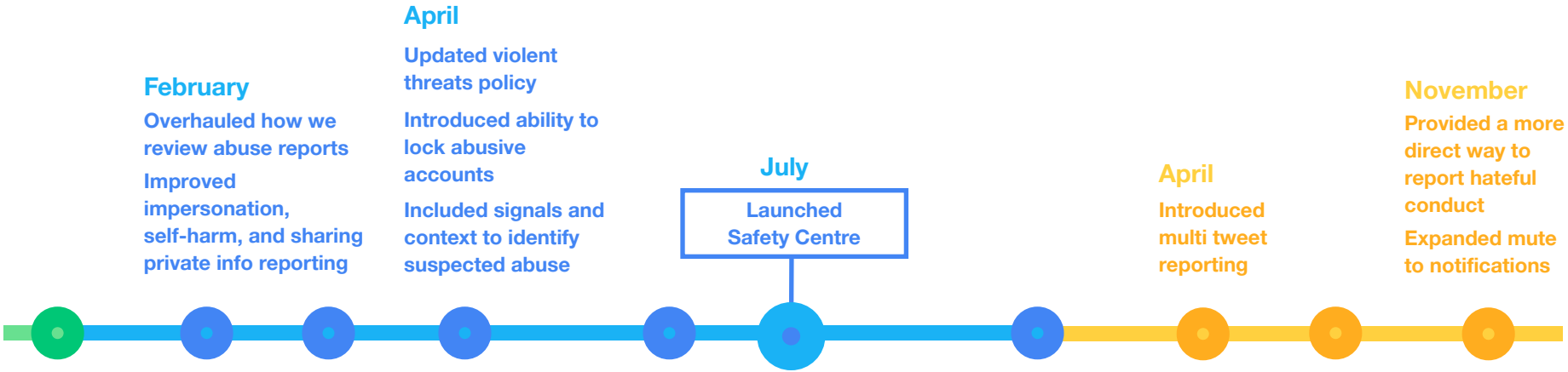
Yours sincerely,

Katy Minshall

Safety Updates

2014 - 2020

Timeline



December
Improved harassment reporting

Enhanced block function

February
Overhauled how we review abuse reports

Improved impersonation, self-harm, and sharing private info reporting

March
Made it easier to report threats to Law Enforcement

April
Updated violent threats policy

Introduced ability to lock abusive accounts

Included signals and context to identify suspected abuse

June
Launched block lists

July
Launched Safety Centre

December
Updated Twitter Rules on abusive behavior and hateful conduct

April
Introduced multi tweet reporting

June
Enabled users to block directly from a tweet

November
Provided a more direct way to report hateful conduct

Expanded mute to notifications

2014 2015 2016


jack
@jack
We're taking a completely new approach to abuse on Twitter. Including having a more open & real-time dialogue about it every step of the way
527
1.4K

Timeline

March

- Leveraged technology to reduce abusive content
- Expanded mute and notification filters
- Began sending in-app report notifications

October

- Updated Non-Consensual Nudity Policy
- Updated Private Information Policy
- Improved communication around appeal process

December

- Implemented account relationship signals to improve witness reporting
- Specifying policy in violation to witnesses who submit reports
- Added new Policy that does not allow hateful display names

February

- Enabled users to report tweets even if blocked
- Stopped the creation of new abusive accounts
- Introduced safe search functionality
- Began collapsing potentially abusive or low-quality tweets

June

- Held first Trust & Safety Council Summit

August

- Updated Safety Center

November

- Sharing offending tweet and explanation for locked accounts
- Updated Twitter Rules on self-harm, spam & related behaviors, and graphic content & adult violence
- Updated Verification Policy on loss of status

2017

- Updated Media Policy to not permit hateful imagery and hate symbols in profile elements
- Began suspending violent groups accounts and removing content that glorifies or condones acts of violence
- Expanded unwanted sexual advances enforcement by integrating relationship interaction signals

Timeline

February

Users can report content that encourages self-harm or suicide

March

Twitter announces Request for Proposal to assist with Health initiative

Users suspended for abusive behaviour emailed with violating content & broken rule

Users can opt-in to get Tweets included in report receipts, in-app and by email

April

Report flow is updated to better define 'protected category'

May

Behaviour-based signals introduced to influence how Tweets appear in Search & Conversation

June

Acquisition of Smyte, a company that specializes in safety, spam, and security issues

Users can avail of security keys for login verification

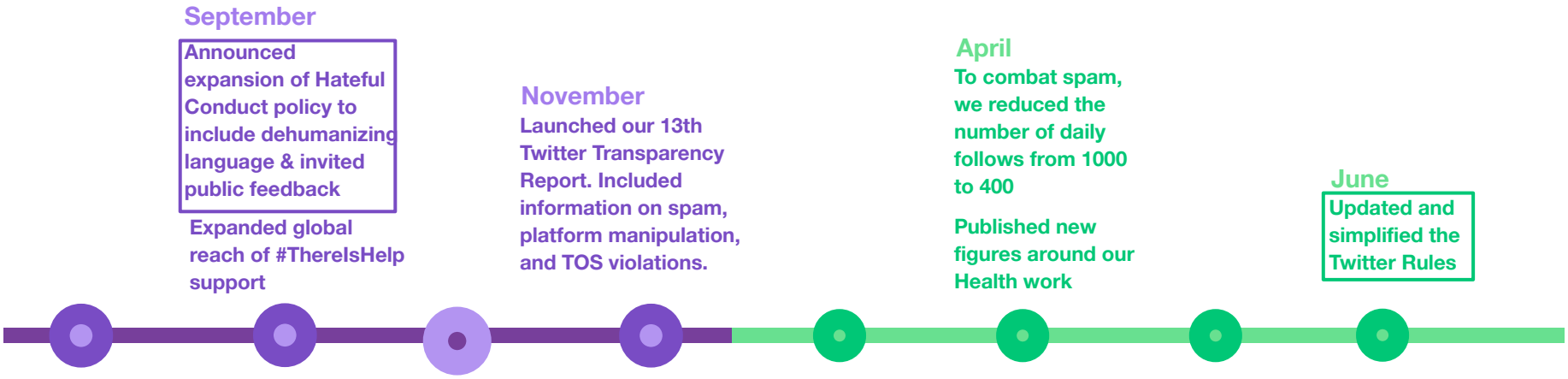
July

More stringent enforcement of the policies related to messages sent during live broadcasts

Twitter to work with Universities of Oxford and Leiden to measure Health work

2018

Timeline



2018

2019

Timeline

June
Introduced a notice to provide more context and clarity around enforcement of Tweets that are in the public interest.

July
Updated our rules against hateful conduct to include language that dehumanizes others on the basis of religion.

September
Announced a new policy to address financial scams on Twitter.

November
Launched our 15th Twitter Transparency Report, which shows a 105% increase in Twitter's proactive enforcement on accounts.

Provided users globally with the ability to moderate replies.

December
Announced we are strengthening our Trust & Safety Council of issue experts to include more diverse voices and foster deeper conversations.

New disclosures on state-backed information operations.

February
Introduced a new policy on synthetic and manipulated media after consulting with civil society and academic experts.

2019

2020