TWITTER SCORECARD:

TRACKING TWITTER'S PROGRESS IN ADDRESSING VIOLENCE AND ABUSE AGAINST WOMEN ONLINE

AMNESTY INTERNATIONAL

Amnesty International is a global movement of more than 7 million people who campaign for a world where human rights are enjoyed by all.

Our vision is for every person to enjoy all the rights enshrined in the Universal Declaration of Human Rights and other international human rights standards.

We are independent of any government, political ideology, economic interest or religion and are funded mainly by our membership and public donations.

First published in 2020 by Amnesty International Ltd
Peter Benenson House, 1 Easton Street, London WC1X 0DW, UK

Index: AMR 51/2993/2020
Original language: English
amnesty.org

Cover photo: © Getty

AMNESTY INTERNATIONAL

INTRODUCTION

Twitter is a social media platform used by hundreds of millions of people around the world to debate, network and share information with each other. As such, it can be a powerful tool for people to make connections and express themselves. But for many women, Twitter is a platform where violence and abuse against them flourishes, often with little accountability.

In 2017, Amnesty International commissioned an online poll of women in 8 countries about their experiences of abuse on social media platforms and used data science to analyze the abuse faced by female Members of Parliament (MPs) on Twitter prior to the UK's 2017 snap election.[1] In March 2018, Amnesty International released *Toxic Twitter: Violence and abuse against women online*, a report exposing the scale, nature and impact of violence and abuse directed towards women in the USA and UK on Twitter.[2] Our research found that the platform had failed to uphold its responsibility to protect women's rights online by failing to adequately investigate and respond to reports of violence and abuse in a transparent manner, leading many women to silence or censor themselves on the platform. While Twitter has made some progress in addressing this issue since 2018, the company continues to fall short on its human rights responsibilities and must do more to protect women's rights online.

Such persistent abuse undermines the right of women to express themselves equally, freely and without fear. As Amnesty International described in *Toxic Twitter*: "Instead of strengthening women's voices, the violence and abuse many women experience on the platform leads women to self-censor what they post, limit their interactions, and even drives women off Twitter completely." Moreover, as highlighted in our research, the abuse experienced is highly intersectional, targeting women of colour, women from ethnic or religious minorities, women belonging to marginalized castes, lesbian, bisexual or transgender women – as well as non-binary individuals – and women with disabilities.

Since the release of *Toxic Twitter* in March 2018, Amnesty International has published a series of other reports – including the *Troll Patrol* study in December 2018, in which Amnesty International and Element AI collaborated to survey millions of tweets received by 778 journalists and politicians from the UK and US throughout 2017 representing a variety of political views, spanning the ideological spectrum.[3] Using cutting-edge data science and machine learning techniques, we were able to provide a quantitative analysis of the unprecedented scale of online abuse against women in the UK and USA.

In November 2019, Amnesty International published research looking at violence and abuse against women on several social media platforms including Twitter in Argentina in the lead up to and during the country's abortion legalization debates.[4] In January 2020, Amnesty International published further research measuring the scale and nature of online abuse faced by women politicians in India during the 2019 General Elections of India.[5] Amnesty International's research detailed further instances of

1. Amnesty International, *Amnesty reveals alarming impact of online violence against women* (Press Release, 20 November 2017), https://www.amnesty.org/en/latest/news/2017/11/amnesty-reveals-alarming-impact-of-online-abuse-against-women/ (last accessed 24 August 2020); also Amnesty Global Insights, *Unsocial Media: Tracking Twitter Abuse against Women MPs* (4 September 2017), https://medium.com/@AmnestyInsights/unsocial-media-tracking-twitter-abuse-against-women-mps-fc28aeca498a (last accessed 24 August 2020)

2. Amnesty International, *Toxic Twitter: A Toxic Place for Women* (Index: ACT 30/8070/2018) March 2018), https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-1/#topanchor (last accessed 24 August 2020)

3. Amnesty International, *Troll Patrol Report* (December 2018), https://decoders.amnesty.org/projects/troll-patrol/findings (last accessed 24 August 2020)

4. Amnesty International, *Corazones Verdes: Violencia online contra las mujeres durante el debate por la legalizacion del aborto en Argentina*. November 2019, https://amnistia.org.ar/corazonesverdes/files/2019/11/corazones_verdes_violencia_online.pdf (last accessed 24 August 2020)

5. Amnesty International, *Troll Patrol India: Exposing the Online Abuse Faced by Women Politicians in India*, 16 January 2020, https://decoders.amnesty.org/projects/troll-patrol-india (last accessed 24 August 2020)

violence and abuse against women on the platform, this time in diverse geographical and linguistic contexts, prompting renewed calls for Twitter to address this urgent and ongoing issue. All of these reports concluded with concrete steps Twitter should take to meet its human rights responsibilities to respect human rights in the context of violence and abuse against women on the platform.

Amnesty International is releasing the Twitter Report Card in an attempt to continue to hold Twitter accountable in protecting women from online violence and abuse on its platform. This Scorecard is designed to track Twitter's global progress in addressing abusive speech against ten indicators, covering **transparency, reporting mechanisms, the abuse report review process, and enhanced privacy and security features**. These indicators were developed based on recommendations that Amnesty International has made in the past regarding how Twitter can best address abusive and problematic content.

WHAT IS VIOLENCE AND ABUSE AGAINST WOMEN ONLINE?

According to the UN Committee on the Elimination of Discrimination against Women, gender-based violence is "violence which is directed against a woman because she is a woman or that affects women disproportionately, and, as such, is a violation of their human rights."[6] The Committee also states that gender-based violence against women includes (but is not limited to) physical, sexual, psychological or economic harm or suffering to women as well as threats of such acts.[7] This may be facilitated by online mediums.

The UN Committee on the Elimination of Discrimination against Women (CEDAW) uses the term 'gender-based violence against women' *to explicitly recognize the gendered causes and impacts of such violence.*[8] The term gender-based violence further strengthens the understanding of such violence as a societal - not individual - problem requiring comprehensive responses. Moreover, CEDAW states that a woman's right to a life free from gender-based violence is indivisible from, and interdependent on, other human rights, including the rights to freedom of expression, participation, assembly and association.[9] According to the Report of the UN Special Rapporteur on violence against women: "The definition of online violence against women therefore extends to any act of gender-based violence against women that is committed, assisted or aggravated in part or fully by the use of ICT, such as mobile phones and smartphones, the Internet, social media platforms or email, against a woman because she is a woman, or affects women disproportionately."[10]

6. UN Women, *General recommendations made by the Committee on the Elimination of Discrimination against Women*, General Recommendation No. 19, 11th session, para. 6., 1992, http://www.un.org/womenwatch/daw/cedaw/recommendations/recomm.htm (last accessed 22 August 2020)

7. Committee on the Elimination of Discrimination against Women, *General recommendation No. 35 on gender-based violence against women, updating general recommendation No. 19*, para. 14, 26 July 2017, CEDAW/C.GC.35, http://tbinternet.ohchr.org/_layouts/treatybodyexternal/Download.aspx?symbolno=CEDAW/C/GC/35&Lang=en (last accessed 22 August 2020)

8. Committee on the Elimination of Discrimination against Women, *General recommendation No. 35 on gender-based violence against women, updating general recommendation No. 19*, 26 July 2017, CEDAW/C.GC.35, http://tbinternet.ohchr.org/_layouts/treatybodyexternal/Download.aspx?symbolno=CEDAW/C/GC/35&Lang=en (last accessed 22 August 2020).

9. Committee on the Elimination of Discrimination against Women, *General recommendation No. 35 on gender-based violence against women, updating general recommendation No. 19*, 26 July 2017, CEDAW/C.GC.35, http://tbinternet.ohchr.org/_layouts/treatybodyexternal/Download.aspx?symbolno=CEDAW/C/GC/35&Lang=en (last accessed 20 August 2020).

10. United Nations Human Rights Council, *Report of the Special Rapporteur on violence against women, its causes and consequences on online violence against women and girls from a human rights perspective*, 18 June – 6 July 2018, A/HRC/38/47, https://undocs.org/pdf?symbol=en/A/HRC/38/47

Violence and abuse against women on social media, including Twitter, includes a variety of experiences such as direct or indirect threats of physical or sexual violence, abuse targeting one or more aspects of a woman's identity (e.g. racism, transphobia, etc.,), targeted harassment, privacy violations such as "doxing" – i.e. uploading private identifying information publicly with the aim to cause alarm or distress, and the sharing of sexual or intimate images of a woman without her consent.[11] Sometimes one or more forms of such violence and abuse will be used together as part of a coordinated attack against an individual which is often referred to as a 'pile-on'. Individuals who engage in a pattern of targeted harassment against a person are often called 'trolls'.[12]

TWITTER'S HUMAN RIGHTS RESPONSIBILITIES

Companies, wherever they operate in the world, have a responsibility to respect all human rights. This is an internationally endorsed standard of expected conduct.[13] The corporate responsibility to respect requires Twitter to take concrete steps to avoid causing or contributing to human rights abuses and to address human rights impacts with which they are involved, including by providing effective remedy for any actual impacts. It also requires them to seek to prevent or mitigate adverse human rights impacts directly linked to their operations or services by their business relationships, even if they have not contributed to those impacts. In practice, this means Twitter should be assessing – on an ongoing and proactive basis – how its policies and practices impact on users' rights to non-discrimination, freedom of expression and opinion, as well other rights, and take steps to mitigate or prevent any possible negative impacts.

As reflected in the Scorecard below, Twitter has made some progress in addressing this issue. They have increased the amount of information available through their Help Center[14] and Transparency Reports,[15] while also launching new public awareness campaigns, expanding the scope of their hateful conduct policy to include language that dehumanizes people based on religion, age, disability or disease, and improving their reporting mechanisms and privacy and security features. These are important steps, and we recognize Twitter's efforts to date. That said, the problem remains, and Twitter must do more in order for women – and all users, in all languages – to be able to use the platform without fear of abuse.

We will update this Scorecard every six months.

11. Amnesty International, *What is online violence and abuse against women?* (20 November 2017), https://www.amnesty.org/en/latest/campaigns/2017/11/what-is-online-violence-and-abuse-against-women/ (last accessed 20 August 2020).

12. Amnesty International, *What is online violence and abuse against women?* (20 November 2017), https://www.amnesty.org/en/latest/campaigns/2017/11/what-is-online-violence-and-abuse-against-women/ (last accessed 20 August 2020).

13. UN Guiding Principles on Business and Human Rights, 2011, http://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf (last accessed 22 August 2020).

14. Twitter, *Help Center*, https://help.twitter.com/en (last accessed 24 August 2020)

15. Twitter, *Twitter Transparency Center*, https://transparency.twitter.com (last accessed 24 August 2020)

DEFINITION OF ABUSIVE AND PROBLEMATIC CONTENT

ABUSIVE CONTENT Tweets that promote violence against or threaten people based on their race, ethnicity, caste, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. Examples include physical or sexual threats, wishes for the physical harm or death, reference to violent events, behaviour that incites fear or repeated slurs, epithets, racist and sexist tropes, or other content that degrades someone.[16]

PROBLEMATIC CONTENT Tweets that contain hurtful or hostile content, especially if repeated to an individual on multiple occasions, but do not necessarily meet the threshold of abuse. Problematic tweets can reinforce negative or harmful stereotypes against a group of individuals (e.g. negative stereotypes about a race or people who follow a certain religion). We believe that such tweets may still have the effect of silencing an individual or groups of individuals. However, we do acknowledge that problematic tweets may be protected expression and would not necessarily be subject to removal from the platform.[17]

METHODOLOGY

This Scorecard synthesizes all of the recommendations we have made to Twitter since 2018 and distills them into ten key recommendations upon which to evaluate the company.[18] These ten recommendations coalesce into four high-level categories: **Transparency, Reporting Mechanisms, Abuse Report Review Process, and Privacy & Security Features**. We have chosen to focus on these four categories of change because of the positive impact we believe each can have on the experiences of women on Twitter. Increasing transparency is the most important step Twitter can take to identify and properly address problems with its handling of abuse on its platform. Making it as easy as possible for users to report abuse and appeal decisions helps Twitter to collaborate directly with its users to make the platform safer. Improving its processes for reviewing reports of abuse enables Twitter to become more efficient at scale while also maintaining higher levels of accuracy and integrity free from bias. Developing more privacy and security features allows Twitter to directly empower its users to protect themselves.

Each individual recommendation is comprised of one to four separate sub-indicators. We then determine whether Twitter has made progress against each sub-indicator, grading each indicator as either ***Not Implemented, Work in Progress***, or ***Implemented***. *Not Implemented* means that Twitter has made no progress to implement our recommendations. *Work in Progress* means that Twitter has made some progress but has not fully implemented our recommendation. *Implemented* means that the company has implemented our recommendation in full. We based our assessment upon a review of two key sources: first, statements made by Twitter in written correspondences with us since 2018; and second, publicly available information on Twitter's website, including its policies, Transparency Reports, blog posts, and Help Center pages. Ahead of publishing the Scorecard, Amnesty International wrote to Twitter to seek an update on the progress of implementing our recommendations and the company's response has been reflected.

16. Amnesty International, *Troll Patrol*, https://decoders.amnesty.org/projects/troll-patrol/findings#abusive_tweet/abusive_sidebar

17. Amnesty International *Troll Patrol*, https://decoders.amnesty.org/projects/troll-patrol/findings#inf_12/problematic_sidebar

18. The Report Card takes into account recommendations Amnesty International has made to Twitter across four reports: Toxic Twitter, Troll Patrol US/UK, Troll Patrol India, and Green Hearts Argentina.

We use sub-indicators to generate a composite score for each recommendation. If Twitter has made no progress against any of the sub-indicators for a specific recommendation, then we grade Twitter as having *Not Implemented* that recommendation. If Twitter has made progress on any of the sub-indicators, then we grade Twitter's efforts for that recommendation as a *Work in Progress*. If Twitter has fully implemented each sub-indicator, then we grade Twitter as having fully *Implemented* that recommendation. If Twitter has made full progress against some sub-indicators but not others, we grade Twitter's effort as a *Work in Progress*. In the context of ongoing public awareness campaigns, we looked at whether these campaigns had addressed all the issues which we raised, as well as whether these campaigns and related materials were available in languages other than English.

A full description of each recommendation and sub-indicator and the reasoning behind our scoring is included below in the section *Detailed Description of Indicators*.

We intend for these scores to be dynamic as Twitter evolves its handling of violence and abuse against women on its platform. We will track Twitter's progress by monitoring Transparency Reports, policy updates, feature launches, and other public announcements, in addition to continuing to engage with Twitter directly.

We would also welcome any further relevant input from civil society organizations and academics working on this issue. If you would like to provide such information, please contact Michael Kleinman, Director of Amnesty International and Amnesty International-USA's Silicon Valley Initiative, at: michael.kleinman@amnesty.org.

© Amnesty International Australia

TWITTER'S SCORECARD IN ADDRESSING VIOLENCE AND ABUSE AGAINST WOMEN ONLINE

CATEGORY	SUBCATEGORY	RECOMMENDATION	SCORE
TRANSPARENCY	Disaggregation	Improve the quality and effectiveness of transparency reports by disaggregating data along types of abuse, geographic region, and verified account status.	WORK IN PROGRESS
	Content Moderators	Increase transparency around the content moderation process by publishing data on the number of moderators employed, the types of trainings required, and the average time it takes for moderators to respond to reports.	NOT IMPLEMENTED
	Appeals	Increase transparency around the appeals process by publishing the volume of appeals received and outcomes of appeals.	NOT IMPLEMENTED
REPORTING MECHANISMS	Feature request	Develop more features to gather and incorporate feedback from users at every stage of the abuse reporting process, from the initial report to the decision.	WORK IN PROGRESS
	Appeals	Improve the appeals process by offering more guidance to users on how the process works and how decisions are made.	IMPLEMENTED
	Public campaign	Continue to educate users on the platform about the harms inflicted upon those who fall victim to abuse through public campaigns and other outreach efforts. This should include sending a notification/ message to users who are found to be in violation of Twitter's rules about the silencing impact and risk of metal health harms caused by sending violence and abuse to another user online.	WORK IN PROGRESS
ABUSE REPORT REVIEW PROCESS	Transparency	Provide clearer examples of what types of behavior rise to the level of violence and abuse and how Twitter assesses penalties for these different types of behavior.	WORK IN PROGRESS
	Automation	Automation should be used in content moderation only with strict safeguards, and always subject to human judgment. As such, Twitter should clearly report out on how it designs and implements automated processes to identify abuse.	NOT IMPLEMENTED
PRIVACY & SECURITY FEATURES	Feature request	Provide tools that make it easier for users to avoid violence and abuse on the platform, including shareable lists of abusive words and other features tailored to the specific types of abuse a user reports.	WORK IN PROGRESS
	Public campaign	Educate users on the platform about the privacy and security features available to them through public campaigns and other outreach channels and make the process for enabling these features as easy as possible.	WORK IN PROGRESS

DETAILED EXPLANATION OF INDICATORS

TRANSPARENCY

1. Improve the quality and effectiveness of transparency reports by disaggregating data along types of abuse, geographic region, and verified account status.

Amnesty International took into account four distinct indicators to assess Twitter's progress:

- Publish the number of reports of abusive or harmful conduct Twitter receives per year. This should include how many of these reports are for directing 'hate against a race, religion, gender, caste or orientation', 'targeted harassment' and 'threatening violence or physical harm'. Twitter should also specifically share these figures for verified accounts on the platform.[19] – **WORK IN PROGRESS**
- Of the disaggregated reports of abuse, publish the number of reports that are found to be – and not be – in breach of Twitter's community guidelines, per year and by category of abuse. Twitter should also specifically share these figures for verified accounts on the platform.[20] – **WORK IN PROGRESS**
- Publish the number of reports of abuse Twitter receives per year that failed to receive any response from the company, disaggregated by the category of abuse reported and by country.[21] – **WORK IN PROGRESS**
- Publish the proportion of users who have made complaints against accounts on the platform and what proportion of users have had complaints made against them on the platform, disaggregated by categories of abuse.[22] – **NOT IMPLEMENTED**

To determine whether Twitter had implemented any of these changes, we reviewed its most recent Transparency Report.[23] We are pleased to see that the most recent Transparency Report – covering the period July through December 2020 – includes more information than previous reports, including total accounts actioned for abuse / harassment and hateful conduct (amongst other categories), the number of reports suspended and the number of pieces of content removed.[24]

That said, the report does not provide data broken down into subcategories of types of abuse, does not distinguish between verified and unverified accounts, does not offer data broken down according to country, does not provide data on how many reports of abuse received no response from the company, and does not provide data on the proportion of users who have made complaints.

In Twitter's response to Amnesty International, it stated that: "While we understand the value and rationale behind country-level data, there are nuances that could be open to misinterpretation, not least that bad actors hide their locations and so can give very misleading impressions of how a problem is manifesting, and individuals located in one country reporting an individual in a different country, which is not clear from aggregate data." Twitter's full response to this report is included as an Annex below.

---

19. Amnesty International, *Toxic Twitter*, Chap. 8; Amnesty International, *Corazones Verdes*, p. 40, 44; Amnesty International, *Troll Patrol India*, p. 49.

20. Amnesty International, *Toxic Twitter*, Chap. 8

21. Amnesty International, *Toxic Twitter*, Chap. 8; Amnesty International, Troll Patrol India, p. 49.

22. Amnesty International, *Toxic Twitter*, Chap. 8

23. Twitter, *Twitter Rules Enforcement*, July to December 2019, https://transparency.twitter.com/en/twitter-rules-enforcement.html (last accessed 25 August 2020)

24. See Twitter India Letter to Amnesty, 29 November 2019 ("At Amnesty's request, transparency report now includes data broken down across a range of key policies detailing the number of reports we receive and the number of accounts we take action on."); Twitter Argentina Letter to Amnesty, Jan 16, 2020.

Although Twitter's response shows some of the considerations at play, Amnesty International is not asking that Twitter provide country-level data about users accused of abuse; instead, we believe Twitter should provide country-level data about users who report abuse, which avoids the issue raised above. Having data on how many users in a given country report abuse, and how this number changes over time, is a critical indicator to help determine whether Twitter's efforts to address this problem are succeeding in a given country. Twitter could also provide contextual information to correct for potential misinterpretation of the data.

They stated in their response letter that, by the time the Score Card comes out, the rules page will be available in other languages - we reflected this in our analysis of Indicator 10 below.

2. Increase transparency around the content moderation process by publishing data on the number of moderators employed, the types of trainings required, and the average time it takes for moderators to respond to reports.

Amnesty International took into account three distinct indicators to assess Twitter's progress:

- Publish the average time it takes for moderators to respond to reports of abuse on the platform, disaggregated by the category of abuse reported. Twitter should also specifically share these figures for verified accounts on the platform.[25] – **NOT IMPLEMENTED**
- Share and publish the number of content moderators Twitter employs, including the number of moderators employed per region and by language.[26] – **NOT IMPLEMENTED**
- Share how moderators are trained to identify gender and other identity-based violence and abuse against users, as well as how moderators are trained about international human rights standards and Twitter's responsibility to respect the rights of users on its platform, including the right for women to express themselves on Twitter freely and without fear of violence and abuse.[27] – **NOT IMPLEMENTED**

To determine whether Twitter had implemented any of these changes, we reviewed its most recent Transparency Report.[28] The report does not include data on the average response time to reports of abuse or the number of content moderators employed broken down by region and language. The report also does not offer any information about the trainings received by content moderators related to gender and identity-based abuse and violence. Other publicly available Twitter pages, such as the Help Center, similarly fail to offer any information about these trainings.

In its response to this report Twitter highlighted the following: "…our strategy is one that combines human moderation capacity with technology. Measuring a company's progress or investment on these important and complex issues with a crude measure of how many people are employed is neither an informative or useful metric. It fails to take into account investments in machine learning, proactive detection, tooling and infrastructure advances...By using new tools to address this conduct from a behavioral perspective, we're able to proactively identify violative accounts and content at scale while reducing the burden on people who use Twitter. We proactively detect 1 in 2 of the Tweets we take

25. Amnesty International, *Toxic Twitter*, Chap. 8; Amnesty International, *Troll Patrol India*, p. 49.

26. Amnesty International, *Toxic Twitter*, Chap. 8; Amnesty International, *Corazones Verdes*, p. 40; Amnesty International, Troll Patrol India, p. 49.

27. Amnesty International, Toxic Twitter, Chap. 8; Amnesty International, *Corazones Verdes*, p. 40, 44; Amnesty International, Troll Patrol India, p. 49.

28. Twitter, *Twitter Rules Enforcement*, July to December 2019, https://transparency.twitter.com/en/twitter-rules-enforcement.html (last accessed 25 August 2020)

down for abuse, compared to one in five Tweets in 2018. This is a significant improvement for those facing abuse, but is not captured by the number of moderators employed."

Amnesty International disagrees with this analysis. The number of content moderators is a critical indicator of Twitter's overall capacity to respond to reports of abusive and problematic content, especially in terms of showing Twitter's capacity – or lack thereof – to cover reports of abuse across different countries and languages, and how this changes over time. Even with investments in machine learning to detect online abuse, it is important to have a measure of the number of human moderators reviewing automated decisions.

The trend towards using machine learning to automate content moderation online also poses risks to human rights. For example, David Kaye, the UN Special Rapporteur on Freedom of Expression, has noted that "automation may provide value for companies assessing huge volumes of user-generated content."[29] He cautions, however, that in subject areas dealing with issues which require an analysis of context, such tools can be less useful, or even problematic, hence the importance of having a sufficient number of human moderators.

3. Increase transparency around the appeals process by publishing the volume of appeals received and outcomes of appeals.

Amnesty International took into account two distinct indicators to assess Twitter's progress:

- Share and publish the number of appeals received for reports of abuse, and the proportion of reports that were overturned in this process, disaggregated by category of abuse.[3] – **NOT IMPLEMENTED**
- Publish information regarding the criteria and decision for granting appeals (or not), year and country-specific number of appeals received, with outcomes.[31] – **NOT IMPLEMENTED**

To determine whether Twitter had implemented any of these changes, we reviewed its most recent Transparency Report, relevant Help Center pages, and various letters.[32] The report does not provide any data on appeals, nor any of the criteria used to make decisions on appeals.

REPORTING MECHANISMS

4. Develop more features to gather and incorporate feedback from users at every stage of the abuse reporting process, from the initial report to the decision.

Amnesty International took into account four distinct indicators to assess Twitter's progress:

- Add an optional question for users who receive a notification about the outcome of any reports on whether or not they were satisfied with Twitter's decision. Twitter should annually share and publish these figures, disaggregated by category of abuse.[33] – **NOT IMPLEMENTED**

---

29. United Nations Human Rights Council, *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*, 6 April 2018, A/HRC/38/35, https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/ContentRegulation.aspx

30. Amnesty International, *Toxic Twitter*, Chap. 8.

31. Amnesty International, *Troll Patrol India*, p. 49.

32. Twitter, *Twitter Rules Enforcement*, July to December 2019, https://transparency.twitter.com/en/twitter-rules-enforcement.html (last accessed 25 August 2020)

33. Amnesty International, *Toxic Twitter*, Chap. 8; Amnesty International, Troll Patrol India, p. 49

- Give users the option to provide a limited character count of context when making reports of violence or abuse to help moderators understand why a report has been made. Twitter should eventually test user satisfaction against reports with an added context and reports without an added context.[34] – **IMPLEMENTED**
- Share information with users who have filed a report of violence and abuse with links and resources for support and suggestions on how to cope with any negative or harmful effects.[35] – **WORK IN PROGRESS**

To determine whether Twitter had implemented any of these changes, we reviewed its most recent Transparency Report,[36] relevant Help Center pages, and various letters it had sent to us over the last two years in response to our requests for updates.

Twitter's Help Center suggests that it provides reporters of abuse with various notifications after they file reports, but Twitter does not request direct feedback from users to assess their satisfaction with the outcome of reports. Even if the platform does somehow collect this data, the information does not appear in the most recent Transparency Report.[37]

In letters Twitter sent to us on 29 November 2019[38] and 16 January 2020,[39] it stated that it had improved its reporting flow by giving users the option to add additional context before submitting a report. The relevant Help Center page confirms that Twitter allows users to flag additional tweets. Twitter also allows users to provide additional context by selecting from a number of pre-selected options (e.g. users are prompted by the question "How is this Tweet abusive or harmful?," and can then select such options as "It's disrespectful or offensive," "Includes private information," "Includes targeted harassment," etc.).[40] In addition, Twitter now provides "in-timeline notice of action taken against reported Tweets." However, Twitter still does not provide a limited character count for users to provide additional context about why they are submitting the report.

In a letter Twitter sent to us on 12 December 2018,[41] it updated us that it now provides "follow-up notifications to individuals that report abuse" and "recommendations for additional actions one can take to improve the experience, such as using the block or mute feature." In another letter it sent to us on 29 November 2019,[42] it updated us that users who report abuse now receive "in-timeline notice of action taken against reported tweets" and no longer see tweets they have reported. While this suggests some progress, we believe Twitter must do more to provide users with links and resources on how to cope with the effects of experiencing violence and abuse on the platform.

In its response to this report, Twitter noted "…while we support the spirit of this proposal and have done so with regards to supporting victims having a single email with the necessary resources to take reports of violent threats to law enforcement, it is unclear how this could be implemented at

34. Amnesty International, *Toxic Twitter*, Chap. 8.

35. Amnesty International, *Toxic Twitter*, Chap. 8.

36. Twitter, *Twitter Rules Enforcement*, July to December 2019, https://transparency.twitter.com/en/twitter-rules-enforcement.html (last accessed 25 August 2020)

37. Twitter, *Twitter Rules Enforcement*, July to December 2019, https://transparency.twitter.com/en/twitter-rules-enforcement.html (last accessed 25 August 2020)

38. Twitter India Letter to Amnesty, 29 November 2019

39. Twitter Argentina Letter to Amnesty, 16 January 2020

40. Twitter, *Report abusive behavior*, https://help.twitter.com/en/safety-and-security/report-abusive-behavior (last accessed 24 August 2020)

41. Twitter US Letter to Amnesty, 12 December 2018

42. Twitter India Letter to Amnesty, 29 November 2019

scale, across all of Twitter's policies. In the case of a single policy alone, there could be a vast range of different issues at hand, with potentially hundreds of relevant partner organisations." Twitter also clarified that its "reporting flow and in-product notifications are translated into 42 main languages."[43]

5. Improve the appeals process by offering more guidance to users on how the process works and how decisions are made.

Amnesty International took into account one distinct indicator to assess Twitter's progress:

- Provide clear guidance to all users on how to appeal any decisions on reports of abuse and clearly stipulate in its policies how this process will work.[44] – **IMPLEMENTED**

A Tweet posted by @TwitterSafety on 2 April 2019 confirms that Twitter has vastly improved its appeals process by launching an in-app appeals process and by improving its response time to appeals requests by 60%. Twitter also confirmed this feature in a letter to us on 29 November 2019.[45] Twitter describes its appeals process on its Help Center, under the heading "Help with Locked or Limited Account."[46]

6. Continue to educate users on the platform about the harms inflicted upon those who fall victim to abuse through public campaigns and other outreach efforts.

Amnesty International took into account two distinct indicators to assess Twitter's progress:

- Create public campaigns and awareness amongst users about the harmful human rights impacts of experiencing violence and abuse on the platform, particularly violence and abuse targeting women and/or marginalized groups. This should include sending a notification/message to users who are found to be in violation of Twitter's rules about the silencing impact and risk of mental health harms caused by sending violence and abuse to another user.[47] – **WORK IN PROGRESS**
- Create public campaigns on Twitter encouraging users to utilize reporting mechanisms on behalf of others experiencing violence and abuse. This can help foster and reiterate Twitter's commitment to ending violence and abuse on the platforms and recognize the emotional burden the reporting process can have on users experiencing the abuse.[48] – **WORK IN PROGRESS**

In November 2019 Twitter launched the Twitter Safety Program campaign. Twitter also recently launched the rules.twitter.com site to provide further information about how it enforces its rules. In its response to this report, Twitter stated: "This new resource is included in emails sent to individuals joining Twitter as well as links to our approach to policy development and enforcement which details factors considered by review teams when determining enforcement actions."

Twitter has also detailed a number of specific campaigns. In one letter dated 29 November 2019, Twitter discussed its efforts to launch a variety of safety-focused campaigns over the years, including the #PositionOfStrength campaign in India in 2016 geared towards women, the #WebWonderWomen

43. Email from Twitter to Amnesty, 25 August 2020
44. Amnesty International, *Toxic Twitter*, Chap. 8.
45. Twitter India Letter to Amnesty, 29 November 2019
46. Twitter, *Help with locked or limited account*, https://help.twitter.com/en/managing-your-account/locked-and-limited-accounts#video (last accessed 27 August 2020)
47. Amnesty International, *Toxic Twitter*, Chap. 8
48. Amnesty International, *Toxic Twitter*, Chap. 8

collaboration also centered on women, the #EduTweet campaign focused on educators and teachers, and "Tweesurfing" aimed at millenials.[49] In another letter dated 16 January 2020, Twitter referred to a recent pact it had signed in Mexico with various stakeholders across academia, civil society, UNESCO, and other international alliances to address gender-based violence in Mexico.[50] In addition, in its response to this report, Twitter stated that it "launched a dedicated gender-based violence search prompt for hotlines and support in local languages in eight Asia Pacific markets: India, Indonesia, Malaysia, Philippines, Thailand, Singapore, South Korea, and Vietnam." Twitter has also posted videos explaining to users who to report problematic content.[51]

These efforts all serve to increase awareness about the harms of abuse and violence on the platform, but we believe Twitter must still do more, particularly in addressing gender-based harms. Specifically, Twitter still has not implemented a feature to notify users who are found to be in violation of Twitter's rules about the silencing impact and risk of mental health harms caused by sending violence and abuse to another user.

Additionally, while this Help Center page provides some guidance on how to help someone you know who is being impacted by online abuse, Twitter should do more to encourage users to report harmful content on behalf of others experiencing violence and abuse, including explicitly encouraging users to report abuse on behalf of others.

ABUSE REPORT REVIEW PROCESS

7. Provide clearer examples of what types of behavior rise to the level of violence and abuse and how Twitter assesses penalties for these different types of behavior.

Amnesty International took into account two distinct indicators to assess Twitter's progress:

- Share specific examples of violence and abuse that Twitter will not tolerate on its platform to both demonstrate and communicate to users how it is putting its policies into practice.[52] – **IMPLEMENTED**
- Share with users how moderators decide the appropriate penalties when accounts users are found to be in violation of the Twitter Rules.[53] – **WORK IN PROGRESS**

To determine whether Twitter had implemented any of these changes, we relied on letters from Twitter, as well as public announcements of recent policy updates.

In a letter dated 29 November 2019, Twitter notified us that it had updated its reporting flow "to offer more detail on what Twitter defines as a 'protected category'," and that it had refreshed the Twitter Rules in June 2019 to simplify them and to add "details such as examples, step-by-step instructions about how to report, and . . . what happens when Twitter takes action."[54] A tweet from @TwitterSafety on June 6, 2019 confirms that this rules refresh took place.

Twitter has also started to provide additional information regarding how moderators decide the appropriate penalties, describing the five factors that moderators take into account. These include:

49. Twitter India Letter to Amnesty, 29 November 2019

50. Twitter Argentina Letter to Amnesty, 16 January 2020

51. Twitter, *How to use Twitter | Reporting Abusive Behavior*, https://www.youtube.com/watch?v=HUEjPiCDaDk (last accessed 24 August 2020)

52. Amnesty International, *Toxic Twitter*, Chap. 8; Amnesty International, Corazones Verdes, p. 44

53. Amnesty International, *Toxic Twitter*, Chap. 8

54. Twitter India Letter to Amnesty, 29 November 2019

"the behavior is directed at an individual, group, or protected category of people; the report has been filed by the target of the abuse or a bystander; the user has a history of violating our policies; the severity of the violation; the content may be a topic of legitimate public interest."[55] That said, Twitter should release more information on how much weight is given to each of these factors, as well as explain how moderators decide between different penalties such as removing the Tweet in question and / or temporarily limiting the user's ability to post new Tweets.

8. Automation should be used in content moderation only with strict safeguards, and always subject to human judgment. As such, Twitter should clearly report out on how they design and implement automated processes to identify abuse.

Amnesty International took into account one distinct indicator to assess Twitter's progress:

- Providing details about any automated processes used to identify online abuse against women, detailing technologies used, accuracy levels, any biases identified in the results and information about how (if) the algorithms are currently on the platform.[56] – **NOT IMPLEMENTED**

To determine whether Twitter had implemented any of these changes, we reviewed Twitter's most recent Transparency Report[57] and other publicly available blogposts and Help Center pages about the use of technology and automation to moderate content. While we found discussions of ways in which Twitter is using technology to take action on problematic content on a larger scale and with greater speed – for example, to combat misinformation during the current COVID-19 pandemic[58] – we did not find any public discussion of the algorithms used or the ways in which Twitter monitors for accuracy and bias, particularly in addressing abuse against women.

In its response to this report Twitter stated it relies on "automated enforcement when the policy violation is of a more serious nature (e.g. child sexual exploitation, violent extremist content)" and where it has assessed it can do so "with high accuracy". It also stated that it does not "permanently suspend accounts based solely on our automated enforcement systems and will continue to look for opportunities to build in human review checks where they are most impactful."

PRIVACY & SECURITY FEATURES

9. Provide tools that make it easier for users to avoid violence and abuse on the platform, including shareable lists of abusive words and other features tailored to the specific types of abuse a user reports.

Amnesty International took into account three distinct indicators to assess Twitter's progress:

- Provide tools that make it easier for women to avoid violence and abuse, such as a list of abusive key words associated with gender or other identity-based profanity or slurs that users can choose from when enabling the filter function. An additional feature could allow users to easily share keywords from their mute lists with other accounts on Twitter.[59] – **WORK IN PROGRESS**

---

55. Twitter, *Our approach to policy development and enforcement policy*, https://help.twitter.com/en/rules-and-policies/enforcement-philosophy#section (last accessed 27 August 2020)

56. Amnesty International, *Troll Patrol India,* p. 49; Amnesty International, Corazones Verdes, p. 33, 44

57. Twitter, *Twitter Rules Enforcement*, July to December 2019, https://transparency.twitter.com/en/twitter-rules-enforcement.html (last accessed 25 August 2020)

58. Twitter India Letter to Amnesty, 29 November 2019 ("More than 50% of tweets actioned on for abuse were surfaced using technology, reducing the burden on those people who may be experiencing abuse and harassment to report to us.")

59. Amnesty International, *Toxic Twitter*, Chap. 8

- Offer personalized information and advice based on personal activity on the platform. For example, share useful tips and guidance on privacy and security settings when users make a report of violence and abuse against them. This should be tailored to the specific category of abuse users report. For example, a person reporting against targeted harassment could be advised how to protect themselves against fake accounts.[60] – **WORK IN PROGRESS**
- Clearly communicate any risks associated with utilizing security features alongside simple ways to mitigate against such risks. For example, if users are taught how to mute notifications from accounts they do not follow - the risk of not knowing about any threats made against them from such accounts should be explained alongside practical ways to mitigate against such risks (e.g. having a friend monitor your Twitter account).[61] – **WORK IN PROGRESS**

To determine whether Twitter had implemented any of these changes, we reviewed letters we received from Twitter, as well as any public announcements of new feature launches.

In addition to its older safety features like blocking and muting accounts, Twitter has launched a variety of new safety features over the last couple of years – including the ability to hide replies to Tweets. However, it has not yet launched the features Amnesty International has proposed in the past, such as shareable lists of keywords associated with gender or other identity-based profanity.

In Twitter's response to this report it notes: "Over the past few years we have expanded people's ability to control their conversations. Aside from Mute and Block, we launched the ability to Hide replies in November 2019 and more recently as of August 2020, we launched new conversation settings that allows people on Twitter, particularly those who have experienced abuse, to choose who can reply to the conversations they start. During the initial experiment we found that these settings prevented an average of three potentially abusive replies while only adding one potentially abusive Retweet with Comment and didn't experience a rise in unwanted Direct Messages. Public research revealed that people who face abuse find these settings helpful."

Twitter has made some progress in personalizing the information it provides to users who report abuse. In a letter to us on 12 December 2018, it reported that it now "provides follow-up notifications to individuals that report abuse, as well as recommendations for additional actions one can take to improve the experience, such as using the block or mute feature."[62] Twitter should go a step further to tailor this advice to the specific category of abuse being reported by the user. For instance, Twitter has partnered with organizations like Glitch, a UK charity campaigning to end online abuse against women and champion digital citizenship, to provide targeted advice to Black Lives Matter activists.[63] These efforts should be expanded.

Twitter also communicates the risks associated with its safety features. According to Twitter's response to this report, they note: "On risks associated with using safety features, we tell people what happens when they use our safety tools including Block, Mute, advanced Mute for words and hashtags, and what happens when individuals are blocked." However, Twitter does not include information or advice on how to mitigate the risks associated with its safety features.

60. Amnesty International, *Toxic Twitter*, Chap. 8
61. Amnesty International, *Toxic Twitter*, Chap. 8
62. Twitter US Letter to Amnesty, 12 December, 2018
63. Twitter UK, https://twitter.com/TwitterUK/status/1277519085014847490?s=20 (last accessed 24 August 2020)

10. Educate users on the platform about the privacy and security features available to them through public campaigns and other outreach channels and make the process for enabling these features as easy as possible.

- Create public campaigns and awareness on Twitter about the different safety features users can enable on the platform. Such campaigns could be promoted to users through various channels such as: promoted posts on Twitter feeds, emails, and in-app notifications encouraging users to learn how to confidently use various safety tools.[64] – **WORK IN PROGRESS**

To determine whether Twitter had implemented any of these changes, we looked at its recent blogposts, tweets, and other public announcements. For example, on 8 November 2019, @TwitterSafety tweeted out a campaign to educate users about features such as blocking, muting, and filtering content. On 5 April 2020, @TwitterSupport tweeted a similar thread.

Twitter noted in its response to this report that it is "continuing to invest in public campaigns and awareness on Twitter about the different safety features." It also explained that in July it concluded "a series of experiments that notify people in-app about our safety tools and launched a notifications quality filter prompt to inform people about this option."

Twitter should continue to run these types of campaigns and expand the channels through which they promote them, including running campaigns in local languages in those countries where abuse against women on the platform is increasing. Twitter should also continue to find new ways to make it as easy as possible for users to enable safety features, including offering these resources in other languages. To that end, Twitter confirmed in an email that the rules.twitter.com page will be published in 17 additional languages in early September, and that the page describing its approach to policy development and enforcement philosophy is currently available in 18 languages.[65]

CONCLUSION

Twitter is still not doing enough to protect women from online violence and abuse.

Since the release of Toxic Twitter in 2018, Amnesty International has continued to highlight the scale of abuse women face on Twitter, including in Argentina, India, the UK and the US. Meanwhile, women have continued to speak out about the abuse they experience on Twitter, and the company's failure to adequately respond.

The persistent abuse women face on the platform undermines their right to express themselves equally, freely and without fear. This abuse is highly intersectional, women from ethnic or religious minorities, marginalized castes, lesbian, bisexual or transgender women - as well as non-binary individuals – and women with disabilities are targets for abuse.

Although the company has made some welcome progress, the Twitter Scorecard shows how much remains to be done. The purpose of the Scorecard is not only to track Twitter's progress, but also to provide concrete recommendations on steps that Twitter should take to address this issue. Of the ten recommendations below, Twitter has, to date, only fully implemented a single one. Using this Scorecard, we will continue to track Twitter's progress on this critical issue going forward.

64. Amnesty International, *Toxic Twitter,* Chap. 8.; Amnesty International, *Corazones Verdes*, p. 44; Amnesty International, *Troll Patrol India*, p. 49.

65. Email from Twitter to Amnesty, 27 August 2020

ANNEX: TWITTER'S RESPONSE

Page 1

Nick Pickles
Global Head of Public Policy
Strategy & Development

Twitter, Inc.
1355 Market St #900
San Francisco, CA 94103

npickles@twitter.com
@nickpickles

26 August 2020

Dear Michael,

Thank you for sharing the findings of your upcoming report and concerns about abuse and violence against women on Twitter. Protecting the health of the public conversation on Twitter is a priority and we continue to invest and make progress in this space.

We appreciate the detailed review citing prior correspondence and acknowledging the investment we have made to protect the health of the conversation. We've made progress in some areas but know we have more to do.

At a high level, we are concerned the Scorecard's framework does not fairly or fully capture our work. A number of items were proposed in regional Amnesty reports, and not in the global report on Twitter, so it is not clear where recommendations are for specific countries or in a global context. If Amnesty is proposing a single scorecard, we would request Amnesty similarly consolidate its recommendations accordingly.

This also has the effect of impacting the scoring where Twitter has fulfilled the request of the initial Amnesty report, but by modifying the recommendations in subsequent regional reports it is now assessed as incomplete.

In your letter, the score is not included for every section, nor is an overall framework of what the score is measured against included, or whether this analysis will extend to other services.

Finally, a number of the suggested approaches are neither relevant or appropriate for Twitter, but these are still scored. We remain concerned that a one-size-fits all approach fails to take into account important distinctions between services.

To your concerns about Twitter's increased reliance on automated content moderation during the pandemic, we rely on automated enforcement when the policy violation is of a more serious nature (e.g. child sexual exploitation, violent extremist content) and have assessed we can do so with high accuracy. We do not permanently suspend accounts based solely on our automated enforcement systems and will continue to look for opportunities to build in human review checks where they are most impactful. This process helps ensure we make the most of

available resources without changing how we evaluate and action on content as a result of COVID-19.

The fall edition of our Twitter Transparency Report (covering the January to June 2020 period) will include a number of improvements in how we define and present enforcement metrics. In the meantime, we will continue to post relevant updates on our policies and metrics in our Coronavirus response blog post.

We responded to a civil society coalition letter on this topic back in July, our response is attached for reference.

Transparency
We believe the Scorecard assessment for items 1.1 and 1.2 and some prescribed indicators are incorrect.

On August 19th we published our most recent Twitter Transparency Report (TTR) within the new Twitter Transparency Center. We now include expanded Rules Enforcement metrics that more closely align with the Twitter Rules including the Hateful Conduct policy, which covers the protected categories listed in your letter. This report expanded the number of policies covered and added more granularity on the actions we take, breaking down the total accounts actioned, the number of accounts suspended and the number of pieces of content removed.

The report card states '*not implemented*' however this data is available in the Transparency Report, which details that in the most recent reporting period July to December 2019:

- Twitter received 4,634,583 reports of hateful conduct, 3,906,683 reports of abuse and 1,722,576 reports of violent threats.
- Twitter took action on 970,109 accounts for violations of our hateful conduct policy, removing 1,445,469 pieces of content and suspending 170,994 accounts. This data is also available relating to our abuse and violent threat policies.

The new Transparency Center includes all our disclosed data in one place and allows for comparison over time. We remain committed to expanding the TTR in future with more granular data, including appeals data. We believe these metrics provide more meaningful transparency and insight into how many accounts were punitively actioned and which policies they violated.

While we have recently updated the Transparency Report, it should be noted that the previous version did include violations broken down across seven key policies (including violent threats and hateful conduct) and the number of reports received.

While we understand the value and rationale behind country-level data, there are nuances that could be open to misinterpretation, not least that bad actors hide their locations and so can give very misleading impressions of how a problem is manifesting, and individuals located in one country reporting an individual in a different country, which is not clear from aggregate data.

On content moderation, we have previously outlined to Amnesty that our strategy is one that combines human moderation capacity with technology. Measuring a company's progress or investment on these important and complex issues with a crude measure of how many people are employed is neither an informative or useful metric. It fails to take into account investments in machine learning, proactive detection, tooling and infrastructure advances, not to mention normalising a narrative that the only way to solve these challenges is to continually hire more people, an approach that risks entrenching an approach that benefits the largest and best resourced companies.

We have teams working around the world to provide timely responses and leverage technology to scale our efforts. Previously, our actions were largely predicated on people reporting accounts or content that violated the Twitter Rules before we could take action. By using new tools to address this conduct from a behavioral perspective, we're able to proactively identify violative accounts and content at scale while reducing the burden on people who use Twitter. We proactively detect 1 in 2 of the Tweets we take down for abuse, compared to one in five Tweets in 2018. This is a significant improvement for those facing abuse, but is not captured by the number of moderators employed.

Similarly, abstract measurements of time may seem useful, but how does that reflect the constant re-prioritisation of reports happening, in part based on the potential severity of harm, or our efforts to limit the impact of bad-faith reporters?

We agree on the need to increase training of moderators on hateful content, particularly identity-based hate, and regularly evaluate the efficacy of content moderation efforts. We will share more on our progress in this area in the future.

With regard to providing individuals with links and resources for support, while we support the spirit of this proposal and have done so with regards to supporting victims having a single email with the necessary resources to take reports of violent threats to law enforcement, it is unclear how this could be implemented at scale, across all of Twitter's policies. In the case of a single policy alone, there could be a vast range of different issues at hand, with potentially

hundreds of relevant partner organisations. We would welcome further discussion on how this could work in practice.

Reporting Mechanisms and Abuse Review Process
We believe the Scorecard assessment for items 6.2 and 7.2 should be revised.

We recently launched a new rules.twitter.com site on how we enforce our rules as part of the Twitter Safety Program campaign launched in November 2019 to educate people about our tools. This new resource is included in emails sent to individuals joining Twitter as well as links to our approach to policy development and enforcement which details factors considered by review teams when determining enforcement actions. When we communicate with people we serve, we do include links to find out more about the process.

Automation
As we have previously discussed, we do not take action for violations of our hateful conduct policy without human review.

Privacy and Security Features
We believe the Scorecard assessment for items 9.1, 9.3, and 10.1 should be revised.

Over the past few years we have expanded people's ability to control their conversations. Aside from Mute and Block, we launched the ability to Hide replies in November 2019 and more recently as of August 2020, we launched new conversation settings that allows people on Twitter, particularly those who have experienced abuse, to choose who can reply to the conversations they start. During the initial experiment we found that these settings prevented an average of three potentially abusive replies while only adding one potentially abusive Retweet with Comment and didn't experience a rise in unwanted Direct Messages. Public research revealed that people who face abuse find these settings helpful.

On risks associated with using safety features, we tell people what happens when they use our safety tools including Block, Mute, advanced Mute for words and hashtags, and what happens when individuals are blocked.

We are continuing to invest in public campaigns and awareness on Twitter about the different safety features. Last month we concluded a series of experiments that notify people in-app about our safety tools and we launched a notifications quality filter prompt to inform people about this option.

There is recent work and policy launches that speak to our commitment to reduce abuse and harassment on Twitter not included in the assessment. Some initiatives include:

- We are testing ways to prompt individuals and add a layer of friction when posting potentially hateful content and sharing articles without having accessed the content first.
- We created and disseminated a resource on the Twitter Rules on Safety and Guidelines on Abuse & Manipulation with best practices for NGOs on account protection and safety tools and will be updating to include the most recent conversation settings launch.
- We launched a dedicated gender-based violence search prompt for hotlines and support in local languages in eight Asia Pacific markets: India, Indonesia, Malaysia, Philippines, Thailand, Singapore, South Korea, and Vietnam.
- We expanded our rules against hateful conduct to include language that dehumanizes others on the basis of religion, age, disability or disease. We plan to expand this policy and are actively consulting with human rights groups to include race, ethnicity, and national origin later in the year.
- Just last month, we updated our URL policy to limit or prevent the spread of URL links to content outside Twitter that promotes violence against, threatens or harasses other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.
- When receiving Direct Messages, we now include the sender's profile information and indicate how the sender is connected to the receiver which can help people quickly identify potentially abusive content.

We respect the work that Amnesty International performs to bring awareness in the field of human rights and support towards vulnerable communities. We welcome further conversations on these issues to learn from your expertise and insights and would be happy to discuss these issues on a call with you and your colleagues.

Best wishes,

Nick Pickles
Global Head of Public Policy Strategy and Development

AMNESTY INTERNATIONAL IS A GLOBAL MOVEMENT FOR HUMAN RIGHTS. WHEN INJUSTICE HAPPENS TO ONE PERSON, IT MATTERS TO US ALL.

CONTACT US

info@amnesty.org

+44 (0)20 7413 5500

JOIN THE CONVERSATION

www.facebook.com/AmnestyGlobal

@AmnestyOnline

Index: AMR 51/2993/2020
September 2020

amnesty.org

AMNESTY INTERNATIONAL