

TWITTER SCORECARD:

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

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.

© Amnesty International 2021

Except where otherwise noted, content in this document is licensed under a Creative Commons (attribution, non-commercial, no derivatives, international 4.0) licence.

<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>

For more information please visit the permissions page on our website:

www.amnesty.org

Where material is attributed to a copyright owner other than Amnesty International this material is not subject to the Creative Commons licence.

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

Index: AFR 53/4722/2021

Original language: English

amnesty.org



Cover illustration: © Amnesty International

**AMNESTY
INTERNATIONAL** 

1. 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 and non-binary persons, Twitter is a platform where violence and abuse against them flourishes, often with little accountability.¹

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 women Members of Parliament (MPs) on Twitter prior to the UK's 2017 snap election.² 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.³ 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 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 and non-binary persons 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 and non-binary persons experience on the platform leads them to self-censor what they post, limit their interactions, and even drives them off Twitter completely." Moreover, as highlighted in our research, the abuse experienced is highly intersectional, targeting women of color, women from ethnic or religious minorities, women belonging to marginalized castes, lesbian, bisexual or transgender women 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.⁴ 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.⁵ In January 2020, Amnesty International published further research measuring the scale and nature of online abuse faced by women politicians in India during

1. Similar abuse also occurs on other platforms including Facebook and Instagram. This report card focuses specifically on Twitter, following Amnesty International's previous research on Twitter, as described below.

2. 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)

3. 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)

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

5. Amnesty International, *Corazones Verdes: Violencia online contra las mujeres durante el debate por la legalización del aborto en Argentina*. November 2019, https://amnistia.org.ar/wp-content/uploads/delightful-downloads/2021/11/Corazones_verdes_Violencia_online.pdf (last accessed 24 August 2020).

the 2019 General Elections of India.⁶ Amnesty International’s research detailed further instances of 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 responsibilities to respect human rights in the context of violence and abuse against women on the platform.

In September 2020, Amnesty International published the first Twitter Scorecard.⁷ This Scorecard was 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.

This is the second edition of the Scorecard, tracking what progress, if any, Twitter has made over the last year as against these ten indicators.⁸ Though Twitter has made some progress, it is far from enough. They have increased the amount of information available through their Help Center⁹ and Transparency Reports,¹⁰ 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; that said, the problem remains. Twitter must do more in order for women and non-binary persons – as well as all users, in all languages – to be able to use the platform without fear of abuse.

6. 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).

7. Amnesty International, Twitter Scorecard (Index: AMR 51/2993/2020, September 2020), <https://www.amnesty.org/en/documents/amr51/2993/2020/en/>

8. In addition to this report focused on the United States, Amnesty International is also simultaneously releasing reports focused on this issue in Argentina and the United States.

9. Twitter, Help Center, <https://help.twitter.com/en> (last accessed 6 July 2021).

10. Twitter, Twitter Transparency Center, <https://transparency.twitter.com> (last accessed 6 July 2021).

2. WHAT IS VIOLENCE AND ABUSE AGAINST WOMEN AND GENDER NON-BINARY PERSONS ONLINE?

According to the UN Committee on the Elimination of Discrimination against Women, gender-based violence includes “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.”¹¹ 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.¹² 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.¹³ 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.¹⁴ 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.”¹⁵

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.¹⁶ 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’.¹⁷

11. 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).

12. 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).

13. 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).

14. 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).

15. 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>.

16. 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).

17. 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).

3. 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.¹⁸ The corporate responsibility to respect human rights 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, freedom of assembly and association, as well other rights, and take steps to mitigate or prevent any possible negative impacts.

4. 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, behavior that incites fear or repeated slurs, epithets, racist and sexist tropes, or other content that degrades someone.¹⁹

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.²⁰

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

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

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

5. VIOLENCE AND ABUSE AGAINST WOMEN AND GENDER NON-BINARY PERSONS ON TWITTER IN SOUTH AFRICA

Twitter has approximately 9.3 million users in South Africa²¹ and is used for various reasons, both positive and negative. On the positive side, Twitter is used to impart accurate information or news, for networking, and to hold people accountable. However, it also has a dark side where it is used to spread disinformation,²² abuse people or incite violence. Recently, it played a key role in inciting deadly riots in South Africa's Gauteng and KwaZulu-Natal provinces²³ and was instrumental in sparking xenophobic violence in the country in 2019.²⁴

For some women users of Twitter in South Africa, threats of violence, abuse and bullying are a common part of their experience of the platform. Amnesty International South Africa found that the abuse women experience on Twitter has forced them to sometimes deactivate their accounts, change the way they interact with the platform, or to self-censor. These experiences are no different to those found in Amnesty International's 2018 Toxic Twitter report, demonstrating the shortcomings of Twitter's policies on abusive and hateful conduct. The violence and abuse experienced by women on Twitter has a "detrimental effect" on the rights of women to express themselves "equally, freely, and without fear."²⁵

Between July and August 2021, Amnesty International South Africa conducted interviews with nine women who live in South Africa and are on Twitter or who had at one point been on Twitter. The interviewees comprised of politicians, journalists, artists, academics, and activists who experienced abuse on the platform. The interviews were semi-structured and focused on the interviewees' experiences of abuse on Twitter, reporting abuse on the platform and Twitter's response to these reports, and whether or not they have seen a decline in the abuse they receive on Twitter over the past two years.

EXPERIENCES OF ABUSE ON TWITTER

The Toxic Twitter Report revealed that "violence and abuse against women on Twitter comes in many forms and targets women in different ways".²⁶ The abuse often contains sexual and/ or misogynistic remarks and may target different aspects of a women's identity, such as their race, gender, or sexuality. The interviewees experienced a range of abusive tweets, some of which threatened violence, rape, or death.

21. SA Social Media Report 2021: "Social Migration", Ornico, 2021, <https://website.ornico.co.za/wp-content/uploads/2021/06/The-SA-Social-Media-Landscape-Report-2021.pdf>.

22. Disinformation is false information shared with the intention to mislead or deceive people. It is different to misinformation, which refers to false information shared without the intention to mislead.

23. DEMOCRACY 2021 PROJECT: *The Dirty Dozen & The Amplification of Incendiary Content During the Outbreak of Unrest in South Africa July 2021*, Centre for Applied Analytics and Behavioral Change, 2021, <https://cabac.org.za/2021/07/30/the-dirty-dozen-the-amplification-of-incendiary-content-during-the-outbreak-of-unrest-in-south-africa-july-2021/>.

24. *HERE BE DRAGONS: Interim Report on Xenophobia on South African Social Media*, Centre for Applied Analytics and Behavioural Change, 2020, https://mcusercontent.com/579fde2866b5cdfda7d96bbb9/files/82b56461-685e-4673-adcd-7358aa34d223/CABC_Interim_Report_on_Xenophobia.pdf.

25. *Toxic Twitter: A Toxic Place for Women* (Index: ACT 30/8070/2018), Amnesty International, March 2018, p. 5, <https://www.amnesty.org/en/latest/news/2018/03/online-violence-against-women-chapter-1/>.

26. *Toxic Twitter: A Toxic Place for Women* (Index: ACT 30/8070/2018), Amnesty International, March 2018, p. 14, <https://www.amnesty.org/en/latest/news/2018/03/online-violence-against-women-chapter-1/>.

Phumzile van Damme, who is a former Member of Parliament (legislator) and a Digital Rights Specialist, became the subject of “trolls”, targeted attacks, and derogatory tropes of women leaders, when she became a public representative. She shared with Amnesty International South Africa some of the death threats that she received and the impact they have had on her.

She said “‘I will find you, and I will stab you’. ‘I will shoot you’, you know, real life threats. And while blocking them [did] kind of get rid of that, you never know if it could actually result in real life harm. And also consistently receiving that abuse does have psychological harm.”²⁷

Legal journalist Karyn Maughan also received similar comments as she was told that “You must be raped and murdered, you must be necklaced”.^{28, 29} Maughan believes that women are more likely to be abused on Twitter.

“With females, for whatever reason, if you are a female in the public space, and you express opinions, you are, I think, disproportionately more likely to be abused,”³⁰ she added.

Another Interviewee, Nomboniso Gasa, who is an adjunct professor, political analyst, cultural critic, feminist and activist, said women who express opinions on Twitter receive sexually charged comments from people who disagree with them. She told us that “if a woman says something, so if I say something about any political figure, people say, ‘Oh, you want to be laid by this person’. So, it has been interesting, this kind of sexualisation of women with whom we disagree. People actually come out and say things like, you know, ‘if you came across me, I would rape you’, or ‘I would not even rape you’.”³¹

Many of our interviewees either observed or experienced the disproportionate number of abusive tweets that Black women receive. Editor in Chief for South Africa’s Eyewitness News, Mahlatse Mahlase, is one of the recipients of tweets that targeted her on the basis of her race and gender. She told us that she has “been insulted as a Black woman in particular, and also portrayed as a Black woman who has no thinking capacity and is just being used by white male counterparts to say what it is. And then there is others around, you know, people threatening your work, rape...”³²

For LGBTQI+ activist, Moude Maodi, the simple act of sharing content of herself with her partner would result in “hate speech, verbal attacks based on sex, based on gender identity, and based on sexual orientation”, as well as threats on her life.

“...they would be mocking our relationship. They would be saying that we will follow you, we will try to find you lesbians who have to be killed. And it was not only against us, you would see that a lot of LGBTI persons even gay friends that we would have, as soon as they share something, then it would be all of that,”³³ said Maodi.

EXPERIENCES OF REPORTING ABUSIVE TWEETS

The experiences interviewees had of reporting abusive tweets on Twitter shows a continued inconsistent enforcement of Twitter rules.

27. Amnesty International South Africa Interview with Phumzile van Damme, 4 August 2021.

28. “Necklacing” is a gruesome violation, notably used during apartheid against individuals accused of betraying the liberation movement. It involves placing a vehicle tyre around an individual’s neck and setting it alight with petrol.

29. Amnesty International South Africa Interview with Karyn Maughan, 10 August 2021.

30. Amnesty International South Africa Interview with Karyn Maughan, 10 August 2021.

31. Amnesty International South Africa Interview with Nomboniso Gasa, 30 June 2021.

32. Amnesty International South Africa Interview with Mahlatse Mahlase, 28 July 2021.

33. Amnesty International South Africa Interview with Moude Maodi, 28 July 2021.

In some instances, Twitter responded quickly to the report, in others, no action was taken. When action was taken, most interviewees shared that it was on tweets that were overtly violent or racist. Phumzile van Damme's experience encapsulates the inconsistency of Twitter's response:

"I used to report a lot of sexism but there is no point in it because it is not going to do anything. It leaves you feeling very kind of defeated to get a response that says, 'there was no violation of our rules'. I mean the only kind of tweets I will report and there will be action is like if it is blatant, like, racist, racism. But I have reported instances where someone said, 'I will shoot you' and there was 'no violation of our rules'. So, you get to a point when you realise that there is actually no point in reporting."³⁴

Another issue that echoed throughout the interviews is anonymity and the use of fake profiles. Interviewees shared that even when Twitter takes action, it is easy for the offender to establish a new account to continue the abuse. Nomboniso Gasa has observed that "sometimes even when they [Twitter] block some people from using the platform, people still go and open other accounts."³⁵

Investigative journalist Pauli van Wyk highlighted that language barrier is sometimes an issue for Twitter. She expanded by saying that "sometimes they [Twitter] are good, you know, sometimes when the threat is really, when it is overtly rape threats, or when it is very clearly in transgression of their regulations and rules they do block. But then it is obviously easy to just create another farce or stupid kind of new [handle]. I do not think they always understand the current situation in our country and what is being said, especially if it is in Xhosa, or Zulu, or Setswana, or Afrikaans. So, if it is in English, the chance of it getting banned is much better."³⁶

Amnesty International continues to call on Twitter to share and publish the number of content moderators it employs per region and by language.

Under the veil of anonymity, users can avoid accountability for their conduct on Twitter. As Rachel Kolisi, the co-founder of the Kolisi Foundation and the wife of South African National Rugby captain, Siya Kolisi, pointed out that "some of what people say on Twitter is illegal in some countries."³⁷

Whilst the South African Constitution protects the right to freedom of expression, that protection does not extend to "the incitement of imminent violence" or "advocacy of hatred that is based on race, ethnicity, gender or religion, and that constitutes incitement to cause harm".³⁸ South Africa is in the process of implementing the Cybercrimes Act, which criminalises the sharing of messages on electronic communication services, like Twitter, that intend to incite violence against a person or group of persons.

The Cybercrimes Act was signed into law earlier this year but will only come into effect when it is given a commencement date.³⁹

THE SILENCING EFFECT

The aim of online violence and abuse against women is to "create a hostile online environment for women with the goal of shaming, intimidating, degrading, belittling or silencing women."⁴⁰ As Amnesty

34. Amnesty International South Africa Interview with Phumzile van Damme, 4 August 2021.

35. Amnesty International South Africa Interview with Nomboniso Gasa, 30 June 2021.

36. Amnesty International South Africa Interview with Pauli van Wyk, 12 July 2021.

37. Amnesty International South Africa Interview, Rachel Kolisi, 1 August, 2021.

38. *The Constitution of the Republic of South Africa*. Section 16 (2), Republic of South Africa, 1996, <https://www.justice.gov.za/legislation/constitution/SACConstitution-web-eng-02.pdf>.

39. *Cybercrimes Act 19 of 2020*, Republic of South Africa, 2020, <https://www.gov.za/documents/cybercrimes-act-19-2020-1-jun-2021-0000>.

40. *Toxic Twitter: A Toxic Place for Women* (Index: ACT 30/8070/2018), Amnesty International, March 2018, p. 22, <https://www.amnesty.org/en/latest/news/2018/03/online-violence-against-women-chapter-1/>.

International found in its Toxic Twitter Report, “Twitter’s inadequate response to violence and abuse against women is leading women to self-censor what they post, limit or change their interactions online, or is driving women off the platform altogether.”⁴¹

For student and youth activist Noor Ahmad, the toxicity that she experienced on Twitter drove her to leave the platform at one stage. She first started using Twitter in 2013 but left and rejoined in 2021. On her return, she has found that the platform has become more toxic for users.

“I left it because it was very toxic. I came back after a couple of years, and literally two weeks in I was like ‘Yeah, this is why I left’,”⁴² she added.

Rachel Kolisi decided to leave Twitter completely after using it for four years. The platform, she told us, became a toxic space for her that negatively impacted her mental health. Karyn Maughan has had to censor what she shares on Twitter after another user with the handle “Death to KM” contacted her sister, claiming to have killed her. Maughan now fears that her engagement on Twitter will negatively impact her family.

She mentioned that “I often have second thoughts about posting stuff with other people mentioned because I am scared they are going to get abused. So, I police myself a lot more than I used to. And that makes me sad. Because I have always got to have a worst-case scenario to link to everything I do.”⁴³ Despite the abuse levelled against her, Mahlatse Mahlase continues to use Twitter because “as a South African, you know the price people have had to pay for you to have freedom of speech and freedom of association.”⁴⁴ However, she has also had to change the way she engages on Twitter.

“I have also decided that it is not a place to share my personal information, it is usually for work content, and so my kind of personal quirky side of life I will share [on] Facebook because I can control [who is] my friend and who is not,”⁴⁵ said Mahlase.

All interviewees, who continue to use the platform, have had to find ways to cope with the scale of abuse they receive.

Poet, musician, and beadwork artist, Ntsiki Mazwai told Amnesty International South Africa that “we should see more Twitter accounts getting shut down when people complain. I definitely think that there should be accountability. But you know, there is none. So, you just learn to adapt.”⁴⁶

THE SITUATION IS NOT IMPROVING

When asked whether they have seen an improvement in the last two years in the number of negative tweets received, the tools that Twitter provides to help screen those tweets, or Twitter’s response when hateful or abusive tweets were reported, all interviewees who could respond said that their experience has worsened. Phumzile van Damme mentioned that “it is gotten worse, because Twitter allows it to get worse, and will continue to get worse”.⁴⁷

41. *Toxic Twitter: A Toxic Place for Women* (Index: ACT 30/8070/2018), Amnesty International, March 2018, p. 46, <https://www.amnesty.org/en/latest/news/2018/03/online-violence-against-women-chapter-1/>.

42. Amnesty International South Africa Interview with Noor Ahmad, 2 July 2021.

43. Amnesty International South Africa Interview with Karyn Maughan, 10 August 2021.

44. Amnesty International South Africa Interview with Mahlatse Mahlase, 28 July 2021.

45. Amnesty International South Africa Interview with Mahlatse Mahlase, 28 July 2021.

46. Amnesty International South Africa Interview with Ntsiki Mazwai, 5 July 2021.

47. Amnesty International South Africa Interview with Phumzile van Damme, 4 August 2021.

Although more mechanisms are available to filter tweets, Karyn Maughan says that her experience of Twitter has not improved.

“A lot of people have complained. There are mechanisms available to you to filter. And you can report and all of this stuff, but in real practical senses, has it improved my experience? No. I kind of have to, oftentimes, just grit my teeth and go into those spaces. And sometimes if I am fragile, or low, like, you know, I just, I did not go there,”⁴⁸ said Maughan.

Nomboniso Gasa said that she thinks that “it is getting worse actually and you can see a real demonstration of troll farms. Whether it is people who are directly paid to do this stuff, or whether it is people who simply are catching up and also want to pile on, you know, I can't quite work it out, but I think it's getting worse.”⁴⁹

Although Ntsiki Mazwai's experience of Twitter has improved over the years, it was a result of personal adjustments that she made, rather than through action from Twitter. She told us that “the changes have been from my side, it is on how I manipulate my tweets, and how I respond, but not necessarily from Twitter.”⁵⁰

“DO NOT CANCEL TWITTER...JUST MAKE IT MORE USER FRIENDLY”⁵¹

For interviewees who remain on the platform, Twitter is a useful tool for their work, businesses, activism and can be a source of positive experiences. However, online violence and abuse, together with inconsistent action from Twitter, sullies their positive experiences of Twitter and causes them to change the way they interact with the platform. The people that Amnesty International South Africa has interviewed, have experienced more abuse on Twitter in the past two years, suggesting the inadequacy of Twitter's policies and the need for stronger action. It is vital for Twitter to uphold its human rights responsibilities and ensure that women are able to engage with the platform “equally, freely and without fear”.

48. Amnesty International South Africa Interview with Karyn Maughan, 10 August 2021.

49. Amnesty International South Africa Interview with Nomboniso Gasa, 30 June 2021.

50. Amnesty International South Africa Interview with Ntsiki Mazwai, 5 July 2021.

51. Amnesty International South Africa Interview with Moude Maodi, 28 July 2021.

6. SCORECARD 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.⁵² 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, Tweets, 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.

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.

52. 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.

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 mental 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.	WORK IN PROGRESS
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

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.⁵³ – **WORK IN PROGRESS**⁵⁴
- 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.⁵⁵ – **WORK IN PROGRESS**⁵⁶
- 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.⁵⁷ – **WORK IN PROGRESS**⁵⁸
- 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.⁵⁹ – **NOT IMPLEMENTED**⁶⁰

To determine whether Twitter had implemented any of these changes, we reviewed its most recent Transparency Report.⁶¹ The most recent Transparency Report – Report 18, covering the period from July to December 2020 – has continued to include the information published in Report 17, including the total accounts actioned for abuse / harassment and hateful conduct (amongst other categories), the number of accounts suspended, and the number of pieces of content removed.⁶² The 18th Transparency Report also includes a few new metrics, such as "impressions", capturing the number of views violative Tweets received prior to removal,⁶³ and new information about the adoption of two-factor authentication.⁶⁴ Twitter is also taking more enforcement action on violative content before it's even

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

54. This indicator is unchanged from the 2020 Twitter Scorecard

55. Amnesty International, Toxic Twitter, Chap. 8.

56. This indicator is unchanged from the 2020 Twitter Scorecard

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

58. This indicator is unchanged from the 2020 *Twitter Scorecard*

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

60. This indicator is unchanged from the 2020 *Twitter Scorecard*

61. Twitter, 18th Transparency Report, July-December 2020, <https://transparency.twitter.com/en/resources.html> (last accessed July 16 2021).

62. 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.

63. "In total, impressions on violative Tweets accounted for less than 0.1% of all impressions for all Tweets globally, from July 1 through December 31. During this time period, Twitter removed 3.8 million Tweets that violated the Twitter Rules; 77% of which received fewer than 100 impressions prior to removal, with an additional 17% receiving between 100 and 1,000 impressions. Only 6% of removed Tweets had more than 1,000 impressions." Twitter Letter to Amnesty, September 27 2021.

64. Twitter, Blog, An update to the Twitter Transparency Center, https://blog.twitter.com/en_us/topics/company/2021/an-update-to-the-twitter-transparency-center (last accessed 16 July 2021).

viewed.⁶⁵ In addition Twitter has recently taken steps to establish a verification process⁶⁶ informed by a consultation process⁶⁷ around the original draft verification policy.⁶⁸

According to the Transparency Report, 82% more accounts were actioned, 9% more accounts suspended, and 132% more content removed compared to the last reporting period. With respect to abuse and harassment, Twitter claims to have deployed more precise machine learning and better detected and took action on violative content, leading to an increase of 142% in accounts actioned compared to the previous reporting period.⁶⁹ Regarding enforcement of the hateful conduct policy (which includes content that incites fear and/or fearful stereotypes about protected categories such as gender, sexual orientation, race, ethnicity, or national origin), 77% more accounts were actioned.

The above information is important and indicative of Twitter's progress to better monitor and report on abusive content online, and report on actions taken to address it. That said, the report still fails to provide data broken down into subcategories of types of abuse; does not offer data broken down on a country level; 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. Twitter doesn't distinguish between verified and unverified accounts, either.

In Twitter's 2020 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."⁷⁰ Twitter's full response to this report is included as an Annex below.

As explained in our previous version of the scorecard, 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, 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.

Twitter stresses that "At Amnesty's request, the 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."⁷¹ However, the transparency report does not provide any data on reported content that failed to receive a response, nor does it provide data on how many reports were reviewed but found to not be in violation of community guidelines. As such, it doesn't specify how many reports were actually reviewed, as opposed to ignored.

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

65. Twitter Letter to Amnesty, September 27 2021.

66. Twitter, Help Center, About Verified Accounts, <https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts> (last accessed 16 July 2021).

67. https://blog.twitter.com/en_us/topics/company/2020/help-us-shape-our-new-approach-to-verification.html

68. <https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts>

69. Twitter, Blog, An update to the Twitter Transparency Center, https://blog.twitter.com/en_us/topics/company/2021/an-update-to-the-twitter-transparency-center (last accessed 16 July 2021).

70. Twitter letter to Amnesty, 26 August 2020

71. Twitter India Letter to Amnesty, Nov 29, 2019; Twitter Arg. Letter to Amnesty, Jan 16, 2020

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.⁷² – **NOT IMPLEMENTED**⁷³
- Share and publish the number of content moderators Twitter employs, including the number of moderators employed per region and by language.⁷⁴ – **NOT IMPLEMENTED**⁷⁵
- 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.⁷⁶ – **NOT IMPLEMENTED**⁷⁷

To determine whether Twitter had implemented any of these changes, Amnesty International reviewed its most recent Transparency Report.⁷⁸ 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 the previous Scorecard report, Twitter argued that "Measuring a company's progress or investment on these important and complex issues with a measure of how many people are employed is neither an informative or useful metric, and only serves to further entrench the largest companies with the greatest resources."⁷⁹ Yet Twitter also acknowledged that their "operations were severely impacted by the Covid-19 pandemic during the latter half of 2020, as was the case with the prior reporting period. Varying country-specific Covid-19 restrictions and adjustments within our teams affected the efficiency of our content moderation work and the speed with which we enforced our policies. We increased our use of machine learning and automation to take a wide range of actions on potentially misleading and manipulative content. Like many organizations – both public and private around the world – the disruptions caused by Covid-19 made an impact on our company and are reflected in some of the data shared today."⁸⁰

Amnesty International firmly believes that 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

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

73. This indicator is unchanged from the 2020 *Twitter Scorecard*

74. Twitter, *Twitter Rules Enforcement*, January to June 2020, <https://transparency.twitter.com/en/twitter-rules-enforcement.html> (last accessed July 6 2021).

75. This indicator is unchanged from the 2020 *Twitter Scorecard*

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

77. This indicator is unchanged from the 2020 *Twitter Scorecard*

78. Twitter, *18th Transparency Report*, July-December 2021, <https://transparency.twitter.com/en/resources.html> (last accessed July 16 2021).

79. Twitter Letter to Amnesty, September 27 2021.

80. Twitter, *Blog*, An update to the Twitter Transparency Center, https://blog.twitter.com/en_us/topics/company/2021/an-update-to-the-twitter-transparency-center (last accessed 16 July 2021).

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. This is especially important during disruptive times such as the Covid-19 pandemic.

The trend towards using machine learning to automate content moderation online also poses risks to human rights. For example, David Kaye, former UN Special Rapporteur on Freedom of Expression, has noted that “automation may provide value for companies assessing huge volumes of user-generated content.”⁸¹ 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. A May 2021 report by the Center for Democracy and Technology further exposes the limitations of algorithmic-driven systems for content moderation.⁸²

Twitter recently admitted that “Many people raised concerns about our ability to enforce our rules fairly and consistently, so we developed a longer, more in-depth training process with our teams to make sure they were better prepared when reviewing reports.”⁸³ Yet no information has yet been shared on how content moderators are trained. This confirms the need to publicly report this information, as it’s otherwise impossible to assess the quality and standards of such training.

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.⁸⁴ –

NOT IMPLEMENTED⁸⁵

- Publish information regarding the criteria and decision for granting appeals (or not), year and country-specific number of appeals received, with outcomes.⁸⁶ – **NOT IMPLEMENTED**⁸⁷

To determine whether Twitter had implemented any of these changes, Amnesty International reviewed its most recent Transparency Report,⁸⁸ relevant Help Center pages, Tweets, and various letters. The report does not provide any data on appeals or any of the criteria used to make decisions on appeals. This is despite Twitter’s assurance that “we remain committed to expanding our future transparency reports with more granular data, including appeals data, and that goal remains a work in progress.”⁸⁹ That said, Twitter has taken steps to improve transparency in the product itself, where users can receive information related to appeals in-app.⁹⁰ Twitter has also started prompting users whether

81. 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>.

82. Center for Democracy and Technology, Dhanaraj Thakur, Emma Llansó, Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis, <https://cdt.org/insights/do-you-see-what-i-see-capabilities-and-limits-of-automated-multimedia-content-analysis/>

83. Twitter, Twitter Safety, Updating our rules against hateful conduct, https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate.html (last accessed July 6 2021).

84. Amnesty International, Toxic Twitter, Chap. 8.

85. This indicator is unchanged from the 2020 Twitter Scorecard

86. Amnesty International, Troll Patrol India, p. 49.

87. This indicator is unchanged from the 2020 *Twitter Scorecard*

88. Twitter, 18th Transparency Report, July-December 2021, <https://transparency.twitter.com/en/resources.html> (last accessed July 16 2021).

89. Twitter letter to Amnesty, September 27, 2021.

90. Twitter letter to Amnesty, September 27, 2021.

they wish to appeal for a sensitive media or misinformation label, as well as implementing in-app suspension banners.⁹¹ However, this information is user-specific, and doesn't provide platform-wide information on an aggregate level.

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 three 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.⁹² – **NOT IMPLEMENTED**⁹³
- 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.⁹⁴ – **IMPLEMENTED**⁹⁵
- 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.⁹⁶ – **WORK IN PROGRESS**⁹⁷

To determine whether Twitter had implemented any of these changes, Amnesty International reviewed its most recent Transparency Report,⁹⁸ 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 reporters of abuse receive notifications after the abuse, although it's not entirely clear what these notifications include beyond "recommendations for additional actions [the user] can take to improve [their] Twitter experience."⁹⁹ Twitter has recently announced that they are exploring the idea of a "Safety Center: A one-stop shop for safety tools. A space where [users] can see the status of [their] reports, blocks, and activity with Twitter Service (even strikes and if [they] are close to being suspended.)"¹⁰⁰

However, the above information is insufficient to determine whether Twitter enables the user to provide direct feedback, or whether this information is personalized to adequately address the user's individual concern. Even if the platform does collect this data, the information does not appear in the most recent Transparency Report.¹⁰¹

91. Twitter letter to Amnesty, September 27, 2021.

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

93. This indicator is unchanged from the 2020 *Twitter Scorecard*

94. Amnesty International, Toxic Twitter, Chap. 8.

95. This indicator is unchanged from the 2020 *Twitter Scorecard*

96. Amnesty International, Toxic Twitter, Chap. 8.

97. This indicator is unchanged from the 2020 *Twitter Scorecard*

98. Twitter, 18th Transparency Report, July-December 2021, <https://transparency.twitter.com/en/resources.html> (last accessed July 16 2021).

99. <https://help.twitter.com/en/safety-and-security/report-abusive-behavior>

100. <https://twitter.com/tapatinah/status/1375224390430961666?s=20>

101. Twitter, Twitter Rules Enforcement, January to June 2020, <https://transparency.twitter.com/en/resources.html> (last accessed July 6 2021)

In letters Twitter sent to us on 29 November 2019 and 16 January 2020, they stated that they had improved their 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.¹⁰² Twitter also allows users to provide additional context by selecting from a number of pre-selected options (e.g. users are prompted by the questions “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”; and “It directs hate against a protected category (e.g., race, religion, gender, orientation, disability etc.).”¹⁰³ In addition, Twitter now provides “in-timeline notice of action taken against reported Tweets.”

A Help Center page also provides additional information on reporting sensitive content.¹⁰⁴ Users are prompted by the question “What issue are you having?” and can then select such options as “An account is harassing me or somebody else”; “An account is directing hate against a protected category, such as race, religion, orientation, sex, disability, or another category”; or “An account is threatening violence or physical harm”,¹⁰⁵ among others.

In a letter Twitter sent to us on 12 December 2018,¹⁰⁶ they updated us that they now provide “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 the letter they sent to us on 29 November 2019,¹⁰⁷ Twitter reported that users no longer see tweets they have reported. The Help Center offers the “Curation style guide”¹⁰⁸ which outlines options for personalizing users’ experience on Twitter. While this points to some progress, we believe that Twitter must do more to provide users with links and resources on how to better cope with the effects of experiencing violence and abuse on the platform.

In their response to the previous 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 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 their “reporting flow and in-product notifications are translated into 42 main languages.”¹⁰⁹

In their response to the current report, Twitter noted: “Improving the experience of reporting is an ongoing effort. As you captured in your letter, we are working on a reporting center and hope to have more to share very soon. We recently relaunched our Help Center in all supported languages to help make it easier for people globally to report content. In the Help Center we also clearly lay out our enforcement options which provide detailed guidance on enforcement and how penalties are assessed.”¹¹⁰ Twitter also noted that they “support organizations that provide assistance to individuals and organizations seeking rapid response emergency help.”¹¹¹

102. Twitter, Help Center, Report abusive behavior, <https://help.twitter.com/en/safety-and-security/report-abusive-behavior> (last accessed July 6 2021).

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

104. Twitter, Help Center, Staying safe on Twitter and sensitive content, <https://help.twitter.com/en/forms/safety-and-sensitive-content/abuse> (last accessed July 6 2021).

105. <https://help.twitter.com/en/forms/safety-and-sensitive-content/abuse>

106. Twitter US Letter to Amnesty, 12 December 2018.

107. Twitter India Letter to Amnesty, 29 November 2019.

108. Twitter, Help Center, Curation style guide, <https://help.twitter.com/en/rules-and-policies/curationstyleguide> (last accessed July 6 2021).

109. Email from Twitter to Amnesty, 25 August 2020.

110. Twitter letter to Amnesty, September 27, 2021.

111. Twitter letter to Amnesty, September 27, 2021.

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.¹¹² – **IMPLEMENTED**¹¹³

Twitter is currently alerting users that “Our support team is experiencing some delays for reviews and responses right now, but we encourage you to report all potential issues.”¹¹⁴ Before the pandemic, however, a Tweet posted by @TwitterSafety on 2 April 2019 confirmed 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 had also confirmed this feature in a letter to us on 29 November 2019.¹¹⁵ Twitter describes their appeals process on their Help Center, under the heading “Help with Locked or Limited Account.”¹¹⁶

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.¹¹⁷ – **WORK IN PROGRESS**¹¹⁸
- 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.¹¹⁹ – **WORK IN PROGRESS**¹²⁰

In November 2019, Twitter launched the Twitter Safety Program campaign. Twitter later launched the rules.twitter.com site to provide further information about how it enforces its rules. In their 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.”

In a 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

112. Amnesty International, Toxic Twitter, Chap. 8.

113. This indicator is unchanged from the 2020 *Twitter Scorecard*

114. <https://help.twitter.com/forms/general>

115. Twitter India Letter to Amnesty, 29 November 2019.

116. Twitter, Help Center, Help with locked or limited account, <https://help.twitter.com/en/managing-your-account/locked-and-limited-accounts> (last accessed July 6 2021).

117. Amnesty International, Toxic Twitter, Chap. 8.

118. This indicator is unchanged from the 2020 *Twitter Scorecard*

119. Amnesty International, Toxic Twitter, Chap. 8.

120. This indicator is unchanged from the 2020 *Twitter Scorecard*

gender-based violence in Mexico.¹²¹ In August 2020, Twitter stated in a subsequent letter that they had “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.”¹²² Twitter has also posted videos explaining to users how to report problematic content.¹²³

An overview of Twitter’s corporate social responsibility-related activities can be found in the 2020 Global Impact Report (published in April 2021).¹²⁴ Of note, Twitter announced that they provided grants to nonprofit partners to “raise awareness to gender-based violence as cases surged during the so-called ‘shadow pandemic’”.¹²⁵ Twitter also partnered with global and local stakeholders in new markets to expand prompts in Twitter’s #ThereisHelp notification service, which now reportedly provides information on gender-based violence in 24 markets.¹²⁶ Similarly, Twitter launched global events on #SaferInternetDay2021, which included safety training and presentations.

While such initiatives specifically recognizing gender-based harm are welcome, they are overall limited to diversity and inclusion efforts, financial support to women-led campaigns or organizations, and discussing issues of gender-based violence in general without acknowledging that such violence is prevalent on the platform itself. These efforts can certainly 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. For instance, Twitter has still 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. In a November 2020 blog post, the Twitter Public Policy team took pride in the fact that “women’s rights have dominated conversations on Twitter [in 2020] with 40 million Tweets so far and counting.”¹²⁷ However, the blog post made no acknowledgment of or reference to gender-based abuse and hate speech on the Twitter platform itself.

Another Twitter Help Center page provides some guidance on how to help someone a user knows who is being impacted by online abuse.¹²⁸ However, 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.

Twitter has also committed to the World Wide Web Foundation’s framework to end online gender-based violence, and stated in their most recent letter that they would have more updates regarding this initiative in the coming months.¹²⁹

121. Twitter Argentina Letter to Amnesty, 16 January 2020.

122. Twitter confirmed this in its letter to Amnesty, 27 September 2021.

123. Twitter, How to use Twitter | Reporting Abusive Behavior, <https://www.youtube.com/watch?v=HUEjPiCDaDk> (last accessed 6 July 2021)

124. <https://about.twitter.com/content/dam/about-twitter/en/company/global-impact-2020.pdf>

125. https://blog.twitter.com/en_us/topics/company/2020/our-work-to-combat-the-shadow-pandemic.html

126. https://blog.twitter.com/en_us/topics/company/2020/our-work-to-combat-the-shadow-pandemic.html

127. witter, Blog, Twitter Public Policy, Our work to combat the ‘shadow pandemic’, https://blog.twitter.com/en_us/topics/company/2020/our-work-to-combat-the-shadow-pandemic.html (last accessed July 6 2021).

128. Twitter, Help Center, How to help someone experiencing online abuse, <https://help.twitter.com/en/safety-and-security/helping-with-online-abuse> (last accessed July 6 2021)

129. Twitter letter to Amnesty, September 27, 2021. For the Web Foundation’s framework, please see webfoundation.org/2021/07/generation-equality-commitments/

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.¹³⁰ – **IMPLEMENTED**¹³¹
- Share with users how moderators decide the appropriate penalties when accounts users are found to be in violation of the Twitter Rules.¹³² – **WORK IN PROGRESS**¹³³

To determine whether Twitter had implemented any of these changes, Amnesty International relied on letters from Twitter, as well as public announcements of recent policy and practices 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.”¹³⁴ A tweet from @TwitterSafety on June 6, 2019 confirms that this rules-refresh took place indeed. In March 2020, Twitter expanded the policy to include age, disability, and disease.¹³⁵ In December 2020, Twitter announced that it had further revised its hate policy to include caste, religion, race, ethnicity, and national origin. However, the blog post didn't include an intersectional analysis on how women of underrepresented castes and religions are disproportionately impacted.¹³⁶

Twitter has also started to provide additional information regarding how moderators decide the appropriate penalties, describing five factors that moderators take into account.¹³⁷ These include the following criteria: “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.”¹³⁸

The Help Center furthermore outlines enforcement options for tweets that are in violation of community guidelines.¹³⁹ These range from limiting Tweet visibility, to requiring Tweet removal and hiding a violating Tweet while awaiting removal.¹⁴⁰ Twitter has rightfully been exploring alternative options to a binary take down/leave up approach to content moderation. Policy enforcement options now include applying a label and/or a warning message to the Tweet; showing a warning to people before they share or like a Tweet; turning off likes, replies, and retweets; and/or providing a link to additional

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

131. This indicator is unchanged from the 2020 *Twitter Scorecard*

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

133. This indicator is unchanged from the 2020 *Twitter Scorecard*

134. Twitter India Letter to Amnesty, 29 November 2019

135. Twitter, Blog, Twitter Safety, Updating our rules against hateful conduct, https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate

136. Twitter, Blog, Twitter Safety, Updating our rules against hateful conduct, https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate.html (last accessed July 6 2021).

137. Twitter, Help Center, Our approach to policy development and enforcement policy, <https://help.twitter.com/en/rules-and-policies/enforcement-philosophy> (last accessed July 6 2021).

138. Twitter, Help Center, Our approach to policy development and enforcement policy, <https://help.twitter.com/en/rules-and-policies/enforcement-philosophy> (last accessed 6 July 2021)

139. Twitter, Help Center, Twitter, Blog, Our range of enforcement options, <https://help.twitter.com/en/rules-and-policies/enforcement-options> (last accessed 6 July 2021).

140. <https://help.twitter.com/en/rules-and-policies/enforcement-options>

information.¹⁴¹ In June 2021, Twitter further clarified that they “do not permit the denial of violent events, including abusive references to specific events where protected categories were the primary victims. This policy now covers targeted and non-targeted content.”¹⁴²

Overall, the Help Center is a platform where users can get valuable information on how to improve their experience on Twitter. Twitter has recently updated the Help Center with additional information in English and is expected to offer translations in other languages. Of note, users can learn more about “how to help someone experiencing online abuse”; “what to do about self-harm and suicide concerns on Twitter”; and how to “report abusive behavior.”¹⁴³ We’re also pleased that Twitter will have more to share soon on their Reporting Center.¹⁴⁴

That said, Twitter should release more information on how much weight is given to the factors outlined above. Twitter should also explain how moderators decide between different penalties. Indeed, the information currently shared is vague and does not provide sufficient details around how moderators evaluate the criteria to adjudicate content. Importantly, there is no gender-specific analysis in the policies around hate speech on the platform.

8. Automation should be used in content moderation only with strict safeguards, and always subject to human judgment. As such, Twitter should clearly report 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.¹⁴⁵ – **WORK IN PROGRESS**¹⁴⁶

To determine whether Twitter had implemented any of these changes, Amnesty International reviewed Twitter’s most recent Transparency Report¹⁴⁷ and other publicly available blogposts, tweets, and Help Center pages related to the use of technology and automation to moderate content.

We found discussions of ways in which Twitter is using algorithmic-driven 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.¹⁴⁸ As mentioned above, Twitter claims to have deployed more precise machine learning, and better detected and took action on abuse and harassment on its platform, leading to an increase of 142% in accounts actioned compared to the previous reporting period.¹⁴⁹ In its most recent letter to Amnesty, Twitter stated that today “65% of the abusive content [they] action is surfaced proactively for human review, instead of relying on reports from people using Twitter.”¹⁵⁰

141. <https://techcrunch.com/2021/07/01/twitter-colorful-misinformation-labels/?guccounter=1>

142. <https://twitter.com/TwitterSafety/status/1399863969246957568?s=20>

143. <https://twitter.com/TwitterSupport/status/1407392249097318402?s=20>

144. Twitter letter to Amnesty, September 27 2021.

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

146. This was updated from “Not Implemented” in the 2020 *Twitter Scorecard*

147. Twitter, 18th Transparency Report, July-December 2021, <https://transparency.twitter.com/en/resources.html> (last accessed July 16 2021).

148. 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.”).

149. Twitter, Blog, An update to the Twitter Transparency Center, https://blog.twitter.com/en_us/topics/company/2021/an-update-to-the-twitter-transparency-center (last accessed 16 July 2021).

150. Twitter Letter to Amnesty, 27 September 2021.

Previously, in its response to the first version of this report, Twitter had 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.”

Twitter recently announced the expansion and growth of their Machine Learning (ML) Ethics, Transparency and Accountability (META) team.¹⁵¹ Currently, publicly available information related to the META work is still very vague. It outlines three goals: researching and understanding the impact of ML decisions; applying learnings to improve Twitter; sharing learnings, and asking for feedback.¹⁵² One publicly available output is the analysis of Twitter’s image cropping algorithm, where the model was tested for gender and race-based biases, and the compatibility with users’ capacity to make their own choices was assessed.¹⁵³

However, Twitter has not yet provided sufficient transparency with respect to algorithmic content moderation. At the time of writing, Twitter has given only limited insight into how it curates content and how users can curate their own timeline in this FAQ.¹⁵⁴ Twitter only shares basic information related to account suggestions and Twitter “moments”.¹⁵⁵ Twitter does not provide much-needed information on datasets or models, nor is there any public discussion of Twitter monitoring efforts for accuracy and bias in addressing abuse against women.

In its most recent letter to Amnesty international, Twitter recognized “the potential risks of automation” and ensured that they will “continue to balance this with safeguards and human review as part of [their] overall strategy.”¹⁵⁶ Twitter also claims they “support the spirit of the Santa Clara Principles on Transparency and Accountability in Content Moderation, and are committed to sharing more detailed information about how [they] enforce the Twitter Rules in future reports.”¹⁵⁷ However, there’s currently no evidence that these principles are implemented internally.

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

151. Twitter also committed to “sharing more results publicly” in its Letter to Amnesty, 27 September 2021.

152. Twitter, Blog, Introducing our Responsible Machine Learning Initiative, https://blog.twitter.com/en_us/topics/company/2021/introducing-responsible-machine-learning-initiative (last accessed July 6 2021).

153. Twitter, Blog, Introducing our Responsible Machine Learning Initiative, https://blog.twitter.com/en_us/topics/company/2021/introducing-responsible-machine-learning-initiative (last accessed July 6 2021).

154. See also Twitter, Help Center, About Twitter’s account suggestions, <https://help.twitter.com/en/using-twitter/account-suggestions> (last accessed on July 6 2021).

155. Twitter, Help Center, Twitter Moments guidelines and principles, <https://help.twitter.com/en/rules-and-policies/twitter-moments-guidelines-and-principles> (last accessed on 6 July 2021).

156. Twitter Letter to Amnesty, 27 September 2021.

157. Twitter, Transparency, Rules Enforcement, <https://transparency.twitter.com/en/reports/rules-enforcement.html#2020-jan-jun> (last accessed on 6 July 2021).

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.¹⁵⁸ – **WORK IN PROGRESS**¹⁵⁹

- 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. 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.¹⁶⁰ – **WORK IN PROGRESS**¹⁶¹
- 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).¹⁶² – **WORK IN PROGRESS**¹⁶³

To determine whether Twitter had implemented any of these changes, Amnesty International 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 or ensuring that communication between Twitter and users is encrypted, Twitter has launched a variety of new safety features over the last couple of years – including the ability to hide replies to Tweets, limit replies¹⁶⁴ such as changing who can reply even after the tweet was sent,¹⁶⁵ turn off the option for people to send emoji reactions and text replies to Fleets via direct message,¹⁶⁶ improve the ability to mute words,¹⁶⁷ remove followers without needing to block an account,¹⁶⁸ new conversation settings that allow users to choose who can reply to the conversations they start,¹⁶⁹ and a “prompt to pause” feature that asks users if they want to review a reply that includes potentially harmful or offensive language before they send it.¹⁷⁰ Twitter is currently rolling out ‘Safety Mode’, which is a feature that temporarily blocks accounts for using potentially harmful language or sending repetitive and uninvited replies or mentions.¹⁷¹ Twitter has also prioritized people from marginalized communities and women journalists when testing Safety Mode, and collaborated with civil society organizations in this product development phase.¹⁷²

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

159. This indicator is unchanged from the 2020 *Twitter Scorecard*

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

161. This indicator is unchanged from the 2020 *Twitter Scorecard*

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

163. This indicator is unchanged from the 2020 *Twitter Scorecard*

164. TechCrunch, *Twitter considers new features for tweeting only to friends, under different personas and more*, <https://techcrunch.com/2021/07/01/twitter-considers-new-features-for-tweeting-only-to-friends-under-different-personas-and-more/?guccounter=1> <https://techcrunch.com/2020/08/11/twitter-now-lets-everyone-limit-replies-to-their-tweets/> (last accessed on 6 July 2021).

165. Twitter, *Twitter Support*, <https://twitter.com/TwitterSafety/status/1415025551773892608?s=20> (last accessed on 16 July 2021).

166. *Twitter Support*, <https://twitter.com/TwitterSupport/status/1370120178919477249?s=20> (last accessed on 6 July 2021). *Twitter Fleets* is a feature that “lets Twitter users post full-screen photos, videos, reaction to tweets or plain text that disappears after 24 hours.” S. Rodriguez, *Twitter to kill Fleets feature, its competitor to Facebook Stories*, CNBC, June 2021, <https://www.cnbc.com/2021/07/14/twitter-to-kill-fleets-feature-its-competitor-to-facebook-snapchat-stories.html#:~:text=Twitter%20introduced%20Fleets%20in%20November,that%20disappears%20after%2024%20hours.>

167. *Twitter Support*, <https://twitter.com/TwitterSupport/status/1407051178689585163?s=20> (last accessed on 6 July 2021).

168. *Twitter Letter to Amnesty*, 27 September 2021.

169. *Twitter Letter to Amnesty*, 27 September 2021.

170. *Twitter Letter to Amnesty*, 27 September 2021.

171. *Twitter Blog*, *Introducing Safety Mode*, 1 September 2021, https://blog.twitter.com/en_us/topics/product/2021/introducing-safety-mode (last accessed 3 September 2021).

172. *Twitter confirmed this in its letter to Amnesty*, 27 September 2021.

In Twitter's response to the previous report, it noted: "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 allow people on Twitter, particularly those who have experienced abuse, to choose who can reply to the conversations they start.¹⁷³ During the initial experiment Amnesty International 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."¹⁷⁴

Twitter has made some progress in personalizing the information it provides to users who report abuse. In a letter sent 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."¹⁷⁵ 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.¹⁷⁶ These efforts should be expanded.

Twitter has also made some progress in improving Twitter access management processes and authentication systems and improving detection and monitoring capabilities. Twitter states that "[s]imilar to how we proactively detect and alert you of suspicious behavior on your account to help you keep it secure, we have internal detection and monitoring tools that help alert us of unusual behavior or possible unauthorized attempts to access our internal tools."¹⁷⁷ Twitter has also committed to invest in privacy and security tools and training for employees and contractors.¹⁷⁸ As of June 2021, users have the option to use security keys as their only form of two-factor authentication (2FA).¹⁷⁹ However none of these features appear to take a gender-sensitive approach.

Recent features also include the new in-app "blue badge" verification application process, which is currently being rolled out. As of 20 May 2021, Twitter provides more clarity as regards the verification eligibility in a new policy, informed by public feedback.¹⁸⁰ Of note, Twitter designers also announced that they're exploring additional features to increase user control and safety. These include prompts that give users the option to revise their reply before it's published if it uses language that could be harmful,¹⁸¹ the ability for users to 'untag' themselves and restrict certain accounts from mentioning them, and other settings to control notifications and prevent further escalation of mass mentions.¹⁸²

173. Twitter confirmed this in its Letter to Amnesty, 27 September 2021.

174. Twitter letter to Amnesty, 26 August 2020.

175. Twitter US Letter to Amnesty, 12 December, 2018.

176. Twitter UK, <https://twitter.com/TwitterUK/status/1277519085014847490?s=20> (last accessed 6 July 2021).

177. Twitter, Blog, Our continued work to keep Twitter secure, https://blog.twitter.com/en_us/topics/company/2020/our-continued-work-to-keep-twitter-secure.html (last accessed on 6 July 2021).

178. Twitter, Blog, Our continued work to keep Twitter secure, https://blog.twitter.com/en_us/topics/company/2020/our-continued-work-to-keep-twitter-secure.html (last accessed on 6 July 2021).

179. Twitter, Blog, Stronger security for your Twitter account, https://blog.twitter.com/en_us/topics/product/2020/stronger-security-for-your-twitter-account (last accessed on 6 July 2021).

180. Twitter, Blog, Relaunching verification and what's next, <https://t.co/t0hksmKns0?amp=1>; and Twitter Support, <https://twitter.com/TwitterSupport/status/1395403954377404417?s=20> (last accessed 6 July 2021).

181. Twitter, Blog, Tweeting with consideration, https://blog.twitter.com/en_us/topics/product/2021/tweeting-with-consideration; and Twitter Support, <https://twitter.com/TwitterSupport/status/1257717113705414658?s=20> (last accessed 6 July 2021). Twitter confirmed this in its letter to Amnesty, 27 September 2021.

182. Twitter, Dominic Camozzi, https://twitter.com/_dcr_/status/1404578211309056006?s=20 (last accessed 6 July 2021).

Twitter has also announced plans to “[run] experiments in the near future to give users more proactive ways to curate their experience, such as providing advanced warning about the tone of a conversation they may be entering, [and explore] new ways for people to leave the conversation by controlling who can @mention them, as well as new ways people can filter out unwanted speech in their replies.”¹⁸³

Finally, Twitter communicates the risks associated with its safety features. In Twitter’s letter of August 2020, 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.”¹⁸⁴ However, Twitter doesn’t clearly lay out risks or consequences of selecting specific options, nor does it suggest targeted actions that users can take to prevent or mitigate these risks.

Unfortunately, despite the above-mentioned features and progress, Twitter has still not launched the features Amnesty International has proposed in the past, such as shareable lists of keywords associated with gender or other identity-based profanity.

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.

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

- 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.¹⁸⁵ – **WORK IN PROGRESS**¹⁸⁶

To determine whether Twitter had implemented any of these changes, Amnesty International looked at its recent blogposts, tweets, and other public announcements. For example, the relevant Help Center page provides an overview of Twitter’s key safety features on the page in short video explanations and tutorials.¹⁸⁷

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 2020 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.” Most recently, Twitter launched the podcast “I Wish I Knew”, co-hosted by researchers at Twitter. The hosts share their research experience, discuss cross-functional collaboration across the company, and explore research-related issues.¹⁸⁸ Twitter also launched a new blog called “Common Thread.”¹⁸⁹

Overall, 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

183. Twitter letter to Amnesty, 27 September 2021.

184. Twitter letter to Amnesty, 26 August 2020.

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

186. This indicator is unchanged from the 2020 *Twitter Scorecard*

187. Twitter, Help Center, How we’re making Twitter safer, <https://help.twitter.com/en/resources/a-safer-twitter> (last accessed 6 July 2021).

188. https://blog.twitter.com/en_us/topics/company/2021/i-wish-i-knew-podcast

189. Twitter Letter to Amnesty, 27 September 2021.

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. As of November 2021, the rules.twitter.com page is available in 42 different languages.¹⁹⁰

Twitter has also developed partnerships with gender justice-oriented organizations. It created the Trust and Safety Council to advise them on content policy-related matters.¹⁹¹ In partnership with UN Women and the UN Human Rights Office, Twitter recently launched custom emojis which would appear alongside localized hashtags, to spread awareness on International Day for the Elimination of Violence Against Women and Human Rights Day.¹⁹² As acknowledged above, Twitter created the hashtag #ThereIsHelp, which suggested helpful information to users searching for certain terms related to domestic and gender-based violence.¹⁹³ They've recently expanded the feature to five additional countries, making them available in a total of 24 markets and 17 languages. While not gender-specific, Twitter has also stated its commitment to fight against anti-Asian racism and xenophobia,¹⁹⁴ and has created resources on best practices for NGOs on account protection and safety tools, among other features. Twitter has also made a commitment to the World Wide Web Foundation's framework to end online gender based violence, as part of the UN Women Generation Equality Forum.¹⁹⁵

However, the above-mentioned campaigns all look at gender-based violence from an external angle, without reflecting on Twitter's own role in facilitating online abuse of women. Importantly, these campaigns fail to educate users on what they can do to prevent or reduce online gender-based violence on Twitter. Twitter has stated its commitment to "increase education and awareness of these types of user tools" and plans "to have more detailed updates over the next months"¹⁹⁶ especially regarding the Safety Playbook they are currently working on for those users who are most often victims of abuse, such as women.¹⁹⁷

190. See <https://help.twitter.com/en/rules-and-policies/twitter-rules>

191. Twitter, Trust and Safety Council, <https://about.twitter.com/en/our-priorities/healthy-conversations/trust-and-safety-council> (last accessed 6 July 2021).

192. Twitter, Blog, Our work to combat the 'shadow pandemic', https://blog.twitter.com/en_us/topics/company/2020/our-work-to-combat-the-shadow-pandemic.html (last accessed 6 July 2021).

193. Twitter, Blog, Our work to combat the 'shadow pandemic', https://blog.twitter.com/en_us/topics/company/2020/our-work-to-combat-the-shadow-pandemic.html (last accessed 6 July 2021).

194. Twitter, Blog, Allyship right now: #StandForAsians, https://blog.twitter.com/en_us/topics/company/2021/allyship-right-now-stand-for-asians.html (last accessed 6 July 2021).

195. Twitter Letter to Amnesty, 27 September 2021.

196. Twitter Letter to Amnesty, 27 September 2021.

197. Twitter Letter to Amnesty, 27 September 2021.

ANNEX: AMNESTY'S LETTER TO TWITTER

Page 1

Nick Pickles, Public Policy Strategy

Cynthia Wong, Legal Director, Human Rights

Twitter, Inc.
1355 Market Street, Suite 900
San Francisco, CA 94103
United States

**AMNESTY
INTERNATIONAL**



AMNESTY INTERNATIONAL
INTERNATIONAL SECRETARIAT
Peter Benenson House, 1 Easton Street
London WC1X 0DW, United Kingdom
T: +44 (0)20 7413 5500 F: +44 (0)20
7956 1157
E: amnestyis@amnesty.org W:
www.amnesty.org

7 September 2021

Dear Nick and Cynthia,

Re: Tracking Twitter's Progress on Addressing Abuse and Violence Against Women

I am writing to provide Twitter an opportunity to respond to the findings of a forthcoming report by Amnesty International.

In March 2018, Amnesty International released [Toxic Twitter](#), exposing experiences of violence and abuse experienced by women on Twitter and failures of the social media platform to uphold its responsibility to protect this group of users.

Such 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 *Toxic Twitter*, the abuse experienced is highly intersectional, touching women of color, women from ethnic or religious minorities, lesbian, bisexual or transgender women – as well as non-binary individuals – and women with disabilities.

Since the release of *Toxic Twitter*, Amnesty International has published a series of other reports – including the [Troll Patrol](#) report in December 2019, measuring violence and abuse against women on Twitter, as well as reports looking at violence and abuse against women on Twitter in [India](#) and [Argentina](#) – detailing further instances of violence and abuse against women on the platform and renewing calls for Twitter to address this urgent and ongoing issue. All of these reports concluded with concrete steps Twitter should take to fulfil its human rights responsibilities moving forward.

Company Registration: 01606776 Registered in England and Wales

In September 2020 Amnesty International published the first [Twitter Scorecard](#). This Scorecard was 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.

According to the 2020 Scorecard, we found that Twitter had made no progress in implementing three of the indicators, had made some progress implementing six of the indicators, and had fully implemented one of the indicators.

As you know, companies, wherever they operate in the world, have a responsibility to respect all human rights. This is an [internationally endorsed standard](#) of expected conduct. 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 taking steps to mitigate or prevent any possible negative impacts.

We are currently preparing the second Twitter Scorecard Card, gauging Twitter's progress in addressing violence and abuse experienced by women on the platform. We have found that, compared to last year, Twitter has made relatively little progress.

Please find attached an Annex that details our analysis and findings. We would welcome any further information from Twitter to help inform this report. We would be grateful to receive your response to these points and to the analysis below by close of business September 21st; if we receive your response at a later date we may not be able to fully reflect it in the report. We will use your response in our report and campaigning materials, including using verbatim quotes. We will also publish your response on our website.

Please respond by email to mkleinman@aiusa.org.

We welcome the opportunity to continue the dialogue with Twitter on these questions.

Yours sincerely,



Michael Kleinman
Director, Silicon Valley Initiative

TWITTER'S RESPONSE

Page 1



27 September 2021

Nick Pickles

Global Head of
Public Policy
Strategy,
Development &
Partnerships
@nickpickles

Twitter, Inc.

1355 Market St #900
San Francisco, CA
94103

Dear Michael,

Thank you for once again sharing the findings of your Twitter Scorecard assessment regarding abuse and violence against women on Twitter.

We continue to pursue our mission to protect the health of the public conversation on Twitter and we've invested considerable resources devoted to this space. As your report highlighted, we believe we have made progress, but know that much of our work continues. We thank you for your detailed review and for this opportunity to provide you with an update on our efforts.

We maintain the belief that a one-size-fits all approach fails to take into account important distinctions between services, while solutions and investments that fall outside of your categories do not translate to an accurate representation of our progress to date. At Twitter we're committed to experimenting in public with product solutions that help address the fundamental problems our users are facing, and empowering them with controls to set their own experience. While many of these changes are not directly captured in your report scorecard, we believe these improvements will ultimately enable our most vulnerable communities to better engage in free expression without fear, a goal we share with Amnesty.

Transparency

On July 14, 2021 we launched our [Twitter Transparency Report 18](#). As noted previously, in our new [Transparency Center](#), we've expanded our Rules Enforcement metrics to include an increased range of policies and a more granular look at the actions we take, breaking down the total accounts actioned, the number of accounts suspended, and the number of pieces of content removed. We believe these metrics provide meaningful transparency and insight into how many accounts were actioned and which policies they violated. The most recent update illustrates that there

was a 77% increase in the number of accounts actioned for violations of our hateful conduct policy.

We're always looking for ways to share more context about our enforcement of the Twitter Rules. As you captured in your letter, we've also added new metrics, including the number of "impressions" or views, violative Tweets received prior to removal, as well as information about the adoption of two-factor authentication. In total, impressions on violative Tweets accounted for less than 0.1% of all impressions for all Tweets globally, from July 1 through December 31. During this time period, Twitter removed 3.8 million Tweets that violated the Twitter Rules; 77% of which received fewer than 100 impressions prior to removal, with an additional 17% receiving between 100 and 1,000 impressions. Only 6% of removed Tweets had more than 1,000 impressions. More broadly, as we work to remove harmful, violative content quickly and at scale, these numbers represent both our present efficiency and where improvement is needed. Our goal is to improve these numbers over time, taking enforcement action on violative content before it's even viewed.

It is important to note that we often action content for a different rule violation than that which was reported, which could lead to some of the asks in your report leading to confusion about the basis of our actions. As we have previously discussed, there is a wider issue about how we could quantify country-level data and how accurate these different calculations would be; for example, individuals can be located in one country and report Tweets sent by someone in a different country. As we stated previously, providing insight into how many accounts were actioned and which policies they violated is a cleaner and more descriptive way of documenting all known instances of abuse on Twitter.

In the area of content moderation, we have previously noted our disagreement with Amnesty's recommendation and outlined 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 measure of how many people are employed is neither an informative or useful metric, and only serves to further entrench the largest companies with the greatest resources.

Regarding appeals data, we remain committed to expanding our future transparency reports with more granular data, including appeals data, and

that goal remains a work in progress. However, in the meantime, we are striving to be more transparent with our users in other formats, such as timely transparency in the product itself. This is a key area we're investing in as we believe it is more valuable to our users to receive this information in-app rather than requiring them to reference our Transparency Report. We continue to experiment with our approach, such as prompting users if they want to appeal for a sensitive media or misinformation label directly in the product. [In-app suspension banners](#) are another way we're communicating with users when they log in, thereby ending reliance on email for these notifications; these banners also provide a link to appeal. Given these workstreams underway, we believe the Scorecard assessment for item 3 should be changed to *Work in Progress*.

Reporting Mechanisms and Abuse Report Process

Improving the experience of reporting is an ongoing effort. As you captured in your letter, we are working on a reporting center and hope to have more to share very soon.

We recently relaunched our Help Center in all [supported languages](#) to help make it easier for people globally to report content. In the Help Center we also clearly lay out [our enforcement options](#) which provide detailed guidance on enforcement and how penalties are assessed.

In addition, we support organizations that provide assistance to individuals and organizations seeking rapid response emergency help. As you noted, we have partnered with health authorities and nonprofit organizations in 27 markets to expand our #ThereIsHelp notification service. When people search terms associated with gender-based violence on Twitter, they will receive a notification with contact information for local hotlines and other resources to encourage them to reach out for help.

This June, as part of the UN Women Generation Equality Forum, we committed to the Web Foundation's framework to end online gender based violence. This pledge was the culmination of a year-long consultation process with over a hundred women focused NGOs to discuss solutions for online gender based violence. Many of the solutions proposed are projects we already have underway, and we are looking forward to sharing more updates in the coming months.

Automation

We appreciate your update to a *Work in Progress* around our automation technology. We continue to step up the level of proactive enforcement across the service and invest in technological solutions to respond to ever-evolving malicious online activity. Today, by using technology, 65% of the abusive content we action is surfaced proactively for human review, instead of relying on reports from people using Twitter.

Although this technology has allowed us to take action on problematic content on a larger scale and with greater speed, we remain aware of the potential risks of automation and continue to balance this with safeguards and human review as part of our overall strategy.

Thank you for highlighting our new ML Ethics, Transparency and Accountability (META) team, and the public results we shared around the testing of our image cropping algorithm for gender and race-based biases. We are excited about the potential impact this team will have on our automation workstreams and we look forward to sharing more results publicly.

Privacy and Security Features

We believe Privacy is about more than protecting data, but also about giving users tools to help create a personalized Twitter experience where they can feel safe.

We have recently added additional privacy and security check-up features such as improving the ability to [remove followers](#) from either a public or protected account without needing to block an account, which can at times result in retaliation.

We also have been exploring features to increase user control and safety. We [launched](#) new conversation settings that allow people on Twitter, particularly those who have experienced abuse, to choose who can reply to the conversations they start.

As you cited, we now have the ability to hide replies to Tweets, and to [change who can reply](#) mid way through a conversation. We are also in the process of testing [Safety Mode](#), a feature that temporarily autoblocks

accounts that use potentially harmful language or send repetitive and uninvited replies or mentions. As a more personalized way to mute and block in real time without the need to share lists, we hope this feature will prove to be a more effective solution by removing the burden on the user. During the product development phase of Safety Mode, we brought in for consultation our civil society Trust and Safety Council, as well as female-identifying journalists and others from marginalized communities, to ensure we had input directly from those most affected when designing and testing this feature.

Another new feature to note is our prompt to pause- when we detect potentially harmful or offensive language in a reply, we may ask people via a [prompt](#) if they want to review it before sending and consider a more considerate reply.

We will also begin running experiments in the near future to give users more proactive ways to curate their experience, such as providing [advanced warning](#) about the tone of a conversation they may be entering. In addition, we are exploring new ways for people to [leave the conversation](#) by controlling who can @mention them, as well as new ways people can filter out unwanted speech in their replies. [Filter and Limit controls](#) are about empowering users to proactively prevent potentially harmful interactions and letting users control the tone of their own conversations. As always we are gathering public feedback on Twitter to help shape our work.

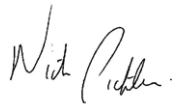
Our next step is to increase education and awareness of these types of user tools, and we plan to have more detailed updates over the next months. We recently collaborated on an initiative with the United Nations Envoy on Youth to publish a [Youth Digital Safety Checklist](#) to help guide users on how to check and safeguard their digital footprint on Twitter. We also look forward to sharing more around a Safety Playbook we are currently working on specifically for those users who are most often victims of abuse, such as women. It will be easily accessible for users to reference and better understand the various tools at their disposal for improving their experience on Twitter in real time.

Finally, we just published a new dedicated space on our blog called [Common Thread](#) where we share stories and interviews about the health of the public conversation on Twitter and the work ahead of us. We hope this

space provides insight into how we think about tough problems and delivers on our goal to bring more transparency to our work.

As always, we appreciate this opportunity to work with you to ensure our platform is serving the needs of its most vulnerable communities and we welcome further dialogue with you on these issues.

Best wishes,


A handwritten signature in black ink that reads "Nick Pickles". The signature is written in a cursive, slightly slanted style.


Nick Pickles
Global Head of Public Policy Strategy, Development and Partnerships



**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](https://twitter.com/AmnestyOnline)

