ARTICLE 19

Twitter Rules and Policies

August 2018

Legal analysis

Executive summary

In this analysis, ARTICLE 19 reviews the compatibility of Twitter's Rules, policies and guidelines (as of August 2018) with international standards on freedom of expression.

The Twitter Rules are complemented by a range of policies on issues such as "hateful conduct," "parody, newsfeed, commentary and fan account" as well as "General guidelines and policies" covering, for instance, the company's policy development and enforcement philosophy ('the Twitter Rules, policies and guidelines'). While the Twitter Rules, policies and guidelines attempt to deal with a wider range of content issues than was previously the case, our analysis shows that they are hard to follow, both in terms of presentation and application. They also generally fall below international standards on freedom of expression, particularly in relation to 'hate speech' and 'terrorism.' Although Twitter's appeals process for the closing of accounts contains a number of positive features, it is unclear whether these policies are consistently applied in practice.

ARTICLE 19 encourages Twitter to bring its Rules, policies and guidelines in line with international human rights law and to continue to provide more information about the way in which those standards are applied in practice.

Summary of recommendations

1. The Twitter Rules, policies and guidelines should be re-organised and consolidated so that the company's rules in relation to particular types of content can be easily found in one place. Consideration should be given to making the Twitter Rules, policies and guidelines available in one consolidated document.

2. Twitter should make clear when the Twitter Rules, policies and guidelines were last updated and identify which parts were amended.

3. Twitter's policies of "hateful conduct" should be clearly presented and should be more closely aligned with international standards on freedom of expression, including by differentiating between different types of prohibited expression on the basis of severity. Importantly, it should provide case studies or more detailed examples of the way in which it applies its "hateful conduct" policies;

4. Twitter should align its definition of terrorism and incitement to terrorism with those recommended by the UN Special Rapporteur on counter-terrorism. In particular, it should avoid the use of vague terms such as "violent extremism", "condone," "celebrate," "glorification" or "promotion" of terrorism;

5. Twitter should give examples of organisations falling within its definition of "violent extremist groups." In particular, it should explain how it complies with various governments' designated lists of terrorist organisations, particularly in circumstances where certain groups designated as 'terrorist' by one government may be considered as legitimate (e.g. freedom fighters) by others. It should also provide case studies explaining how it applies its standards in practice;

6. Twitter should explain in more detail the relationship between "threats," "harassment," and "online abuse" and distinguish these from "offensive content" (which should not be limited as such). Further, Twitter should provide detailed examples or case studies of the way in which it applies its standards in practice, including with a view to ensuring protections for minority and vulnerable groups;

7. Twitter should state more clearly that offensive content will not be taken down as a matter of principle unless it violates other rules;

8. Twitter should make more explicit reference to the need to balance the protection of the right to privacy with the right to freedom of expression. In so doing, it should refer to the criteria developed, *inter alia,* in the Global Principles on the Protection of Freedom of Expression and Privacy;

9. Twitter should be more transparent and explain in more detail how its algorithms detect 'fake' accounts, including by listing the criteria on the basis of which these algorithms operate;

10. Twitter should explain how the measures it is adopting to fight fake accounts, bots etc. are different from removing false information.

11. Twitter should also define or further define what it considers to be "suspicious activity," "bad actors" or "platform manipulation."

12. Twitter should ensure that its appeals process complies with the Manila Principles on Intermediary Liability, particularly as regards notice, the giving of reasons, and appeals processes;

13. Twitter should be more transparent about its use of algorithms to detect various types of content, such as 'terrorist' videos, 'fake' accounts or 'hate speech;'

14. Twitter should clarify whether it relies on a trusted flagger system, and if so it should provide information about its system, including identifying members of the scheme and the criteria being applied to join it;

15. Twitter should provide case studies of the way in which it applies its sanctions policy;

16. Twitter should provide disaggregated data on the types of sanctions it applies in its Transparency Report;

Table of contents

Introduction

Since its inception in 2006, Twitter has grown into a multi-million-dollar company, with an average of 335 million monthly active users. It is now a critical gateway for the exercise of freedom of expression online, allowing the rapid exchange of news, information, opinions and ideas on a massive scale. By the same token, it has also become a major player in the regulation and moderation of online content.

Over the years, Twitter has amended its rules on content to deal with major controversies about its perceived permissive approach to online abuse and trolls. Since 2016, for instance, Twitter has:

- Developed and expanded its "hateful conduct", media and enforcement policies to include abusive usernames and hateful imagery;
- Expanded the list of abusive behaviours that are prohibited on the platform to include "unwanted sexual advances, posting or sharing intimate photos or videos of someone that were produced or distributed without their consent, wishes or hopes of harm, and threats to expose or hack someone;"
- Updated its rules around violence and physical harm to include the glorification of violence and violent extremist groups.

ARTICLE 19 welcomes Twitter's efforts in seeking to clarify its rules on content moderation and the way in which it approaches enforcement of its policies. The latest Twitter Transparency Report, which features content removals on the basis of the company's Terms of Service is another welcome development.

However, we find that the Twitter Rules and related policies and guidelines remain difficult to understand and often fall short of international standards on freedom of expression. In particular, Twitter imposes restrictions on "violent extremism" that are inconsistent with applicable international standards in this area. Most rules also remain very broad in scope, leaving significant discretion to Twitter in their implementation. As such, they are highly likely to lead to inconsistent application. This is all the more so given that Twitter does not provide case studies or detailed examples of its internal 'case-law.'

ARTICLE 19 believes that social media companies, including Twitter, should respect international standards on human rights consistent with the UN Guiding Principles on Business and Human Rights (the UN Guiding Principles). Although these companies are not subjects of international law *per se*, they have human rights responsibilities as central enablers of freedom of expression online. This is especially the case for companies such as Twitter, which occupy such a prominent position in the Internet ecosystem.

In this analysis, we first set out the international standards on freedom of expression that companies should respect, consistent with the UN Guiding Principles. We then analyse the Twitter Rules and related policies and guidelines [1] in some key areas, focusing on 'hate speech,' 'terrorist' content, privacy and morality-based restrictions on content, and 'fake news.' We also examine Twitter's content removal processes and sanctions. Each section contains recommendations on how to bring the Twitter Rules, policies and guidelines in line with international standards on freedom of expression.

[1] As noted earlier, this analysis reviews the Twitter Rules as of August 2018. It appears that this corresponds to the November 2017 version, though this could not be ascertained; see, e.g. BBC, Twitter's Rewritten Rules Published, 3 November 2017; or Twitter, The Twitter Rules: A Living Document, 7 August 2018.

International human rights standards

ARTICLE 19's comments on Twitter Rules are informed by international human rights law and standards.

The right to freedom of expression

The right to freedom of expression is protected by Article 19 of the Universal Declaration of Human Rights (UDHR),[2] and given legal force through Article 19 of the International Covenant on Civil and Political Rights (ICCPR).

The scope of the right to freedom of expression is broad. It requires States to guarantee all people the freedom to seek, receive or impart information or ideas of any kind, regardless of frontiers, through any media of a person's choice. The UN Human Rights Committee (HR Committee), the treaty body of independent experts monitoring States' compliance with the ICCPR, has affirmed that the scope of the right extends to the expression of opinions and ideas that others may find deeply offensive.[3]

While the right to freedom of expression is fundamental, it is not absolute. A State may, exceptionally, limit the right under Article 19(3) of the ICCPR, provided that the limitation is:

- **Provided for by law**, i.e. any law or regulation must be formulated with sufficient precision to enable individuals to regulate their conduct accordingly;
- **In pursuit of a legitimate aim**, listed exhaustively as: respect of the rights or reputations of others; or the protection of national security or of public order (*ordre public*), or of public health or morals;
- **Necessary and proportionate in a democratic society**, i.e. if a less intrusive measure is capable of achieving the same purpose as a more restrictive one, the least restrictive measure must be applied.[4]

Further, Article 20(2) ICCPR provides that any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence must be prohibited by law.

The same principles apply to electronic forms of communication or expression disseminated over the Internet.[5]

Social media companies and freedom of expression

International bodies have also commented on the relationship between freedom of expression and social media companies in several areas.

[2] Although the UDHR (adopted as a resolution of the UN General Assembly) is not strictly binding on states, many of its provisions are regarded as having acquired legal force as customary international law since its adoption in 1948; see *Filartiga v. Pena-Irala*, 630 F. 2d 876 (1980) (US Circuit Court of Appeals, 2nd circuit).
[3] See HR Committee, General Comment No. 34 on Article 19: Freedoms of opinion and expression, CCPR/C/GC/34, 12 September 2011, para 11
[4] HR Committee, *Belichkin v. Belarus*, Comm. No. 1022/2001, U.N. Doc. CCPR/C/85/D/1022/2001 (2005).
[5] General Comment No. 34, *op.cit.*, para 43

Intermediary liability

The four special mandates on freedom of expression have recognised for some time that immunity from liability is the most effective way of protecting freedom of expression online. For example, in their 2011 Joint Declaration, they recommended that intermediaries should not be liable for content produced by others when providing technical services, and that liability should only be incurred if the intermediary has specifically intervened in the content, which is published online.[6]

In 2011 the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (Special Rapporteur on FOE) stated that censorship should never be delegated to a private entity, and that States should not use or force intermediaries to undertake censorship on their behalf. [7] He also noted that notice-and-takedown regimes – whereby intermediaries are encouraged to takedown allegedly illegal content upon notice lest they be held liable – were subject to abuse by both States and private actors; and that the lack of transparency in relation to decision-making by intermediaries often obscured discriminatory practices or political pressure affecting the companies' decisions.[8]

In 2018, the UN Special Rapporteur reiterated that States should refrain from imposing disproportionate sanctions, whether heavy fines or imprisonment, on Internet intermediaries, given their significant chilling effect on freedom of expression.[9] Furthermore, the Special Rapporteur recommended that States should publish detailed transparency reports on all content-related requests issued to intermediaries and involve civil society organisations in all regulatory considerations.[10]

Human rights responsibilities of the private sector

There is a growing body of recommendations from international and regional human bodies that social media companies have a responsibility to respect human rights.

- The **UN Guiding Principles** provide a starting point for articulating the role of the private sector in protecting human rights on the Internet.[11] They recognise the responsibility of business enterprises to respect human rights, independent of State obligations or the implementation of those obligations. In particular, they recommend that companies should:[12]
  - Make a public statement of their commitment to respect human rights, endorsed by senior or executive-level management;
  - Conduct due diligence and human rights impact assessments in order to identify, prevent, and mitigate against any potential negative human rights impacts of their operations;
  - Incorporate human rights safeguards by design in order to mitigate adverse impacts, and build leverage and act collectively in order to strengthen their power vis-a-vis government authorities;

[6] The Joint Declaration on Freedom of Expression and the Internet (the 2011 Joint Declaration), adopted by the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (Special Rapporteur on FOE), the OSCE Representative on Freedom of the Media, the OAS Special Rapporteur on Freedom of Expression and the ACHPR Special Rapporteur on Freedom of Expression and Access to Information, 1 June 2011.
[7] The Report of the Special Rapporteur on FOE, 16 May 2011, A/HRC/17/27, para 43
[8] *Ibid.,* para 42
[9] Report of the Special Rapporteur on FOE, 6 April 2018, A/HRC/38/35, para. 66.
[10] *Ibid.*, para. 69.
[11] Guiding Principles on Business and Human Rights: Implementing the UN 'Protect, Respect and Remedy' Framework, developed by the Special Representative of the Secretary-General on the issue of human rights and transnational corporations and other business enterprises, report of the Special Representative of the Secretary-General on the Issue of Human Rights and Transnational Corporations and Other Business Enterprises, John Ruggie, 7 April 2008, A/HRC/8/5A/HRC/17/31. The Human Rights Council endorsed the Guiding Principles in its resolution 17/4 of 16 June 2011.
[12] *Ibid.*, Principle 15

- Track and communicate performance, risks and government demands; and
- Make remedies available where adverse human rights impacts are created.

- In his **May 2011 report to the United Nations Human Rights Council** (Human Rights Council), the Special Rapporteur on FOE highlighted that, while States are the duty-bearers of human rights, Internet intermediaries also have a responsibility to respect human rights, and referenced the UN Guiding Principles in this regard.[13] The Special Rapporteur also noted the usefulness of multi-stakeholder initiatives, such as the Global Network Initiative (GNI), which encourage companies to undertake human rights impact assessments of their decisions as well as to produce transparency reports when confronted with situations that may undermine the rights to freedom of expression and privacy.[14] He further recommended that, *inter alia,* intermediaries should only implement restrictions to these rights after judicial intervention; be transparent in respect of the restrictive measures they undertake; provide, if possible, forewarning to users before implementing restrictive measures; and provide effective remedies for affected users.[15] The Special Rapporteur on FOE also encouraged corporations to establish clear and unambiguous terms of service in line with international human rights norms and principles, and; to continuously review the impact of their services on the freedom of expression of their users, as well as the potential pitfalls of their misuse.[16]

- In his **June 2016 Report to the Human Rights Council**,[17] the Special Rapporteur on FOE additionally enjoined States not to require or otherwise pressure the private sector to take steps that unnecessarily or disproportionately interfere with freedom of expression, whether through laws, policies, or extra-legal means. He further recognised that "private intermediaries are typically ill-equipped to make determinations of content illegality"[18] and reiterated criticism of notice-and-takedown frameworks for "incentivising questionable claims and for failing to provide adequate protection for the intermediaries that seek to apply fair and human rights-sensitive standards to content regulation."[19]

- In his **April 2018 Report to the Human Rights Council**,[20] the Special Rapporteur on FOE urged social media companies to recognise human rights law as the authoritative global standard for ensuring FOE on their platforms, and to design and implement their content regulation policies accordingly, rather than allowing their policies to depend on the varying national laws of States or their own commercial interests.[21] As such, he called on social media companies to ensure that content-related actions at all stages of their operations, from rule-making to implementation, are guided by the same standards of legality, necessity, proportionality and non-discrimination that bind State regulation of expression.[22] Furthermore, the Special Rapporteur stressed the need that social media companies engage more actively with civil society organisations[23] and open themselves up to public

[13] The May 2011 Report of the Special Rapporteur on FOE, *op.cit.*, para 45
[14] *Ibid.* para 46
[15] *Ibid.*, paras 47 and 76
[16] *Ibid.*, paras 48 and 77
[17] Report of the Special Rapporteur on FOE, 11 May 2016, A/HRC/32/38; para 40-44
[18] *Ibid*
[19] *Ibid.*, para 43
[20] Report of the Special Rapporteur on FOE, 6 April 2018, A/HRC/38/35.
[21] *Ibid.*, paras. 41-43.
[22] *Ibid.*, paras. 45-48. According to the Special Rapporteur, "[c]ompanies committed to implementing human rights standards throughout their operations — and not merely when it aligns with their interests — will stand on firmer ground when they seek to hold States accountable to the same standards. Furthermore, when companies align their terms of service more closely with human rights law, States will find it harder to exploit them to censor content."
[23] *Ibid.*, para. 54.

accountability mechanisms (such as a social media council)[24], in order to achieve higher levels of transparency and consistency in content moderation.

- In his **2013 Report, the Special Rapporteur for Freedom of Expression of the Inter-American Commission on Human Rights** (OAS Special Rapporteur on FOE), also noted the relevance of the UN Guiding Principles[25] and further recommended, *inter alia,* that private actors establish and implement service conditions that are transparent, clear, accessible, and consistent with international human rights standards and principles, and ensure that restrictions derived from the application of the terms of service do not unlawfully or disproportionately restrict the right to freedom of expression.[26] He also encouraged companies to publish transparency reports about government requests for user data or content removal;[27] challenge requests for content removal or requests for user data that may violate the law or internationally recognised human rights;[28] notify individuals affected by any measure restricting their freedom of expression and provide them with non-judicial remedies;[29] and take proactive protective measures to develop good business practices consistent with respect for human rights.[30]

- In the **2016 report on Standards for a Free, Open and Inclusive Internet,**[31] the OAS Special Rapporteur on FOE recommended that, *inter alia,* companies make a formal and high-level commitment to respect human rights, and back up this commitment with concrete internal measures and systems; seek to ensure that any restriction based on companies' terms of service do not unlawfully or disproportionately restrict freedom of expression; and put in place effective systems of monitoring, impact assessments, and accessible, effective complaints mechanisms.[32] He also highlighted the need for companies' policies, operating procedures and practices to be transparent.[33]

- At the European level, in an **issue paper on the rule of law on the Internet and in the wider digital world**, the Council of Europe Commissioner for Human Rights recommended that States stop relying on private companies which control the Internet to impose restrictions that violate States' human rights obligations.[34] He recommended that further guidance should be developed on the responsibilities of businesses in relation to their activities on (or affecting) the Internet, in particular to cover situations in which companies may be faced with demands from governments that may be in violation of international human rights law.[35]

- Similarly, in its **Recommendation on the protection of human rights with regard to social networking services**, the Committee of Ministers of the Council of Europe, recommended that social media companies should respect human rights and the rule of law, including procedural safeguards.[36] Moreover, in its March 2018 **Recommendation on the roles and**

[24] *Ibid.*, para. 58.
[25] OAS Special Rapporteur on FOE, Freedom of Expression and the Internet, 2013, paras 110-116. It notes that "the adoption of voluntary measures by intermediaries that restrict the freedom of expression of the users of their services - for example, by moderating user-generated content - can only be considered legitimate when those restrictions do not arbitrarily hinder or impede a person's opportunity for expression on the Internet".
[26] *Ibid.*, paras 111-112
[27] *Ibid.*, para 113
[28] *Ibid.*, para 114
[29] *Ibid.*, para 115
[30] *Ibid.*, para 116
[31] OAS Special Rapporteur on FOE, Standards for a Free, Open and Inclusive Internet, 2016, paras 95-101
[32] *Ibid.*, para 98
[33] *Ibid.*, para 99
[34] The rule of law on the Internet and in the wider digital world, Issue paper published by the Council of Europe Commissioner for Human Rights, CommDH/IssuePaper (2014) 1, 8 December 2014
[35] *Ibid.*, p. 24
[36] Committee of Ministers of Council of Europe, Recommendation CM/Rec (2012)4 of the Committee of Ministers to Member States on the protection of human rights with regard to social networking services, adopted by the Committee of Ministers on 4 April 2012 at the 1139th meeting of the Ministers' Deputies. These recommendations

responsibilities of internet intermediaries, the Committee of Ministers adopted detailed recommendations on the responsibilities of Internet intermediaries to protect the rights to freedom of expression and privacy and to respect the rule of law.[37] It recommended that companies should be transparent about their use of automated data processing techniques, including the operation of algorithms.

Additionally, recommendations that social media companies should respect international human rights standards have been made by a number of civil society initiatives.

- The **Manila Principles on Intermediary Liability** elaborate the types of measures that companies should take in order to respect human rights.[38] In particular, they make clear that companies' content restriction practices must comply with the tests of necessity and proportionality under human rights law,[39] and that intermediaries should provide users with complaints mechanisms to review decisions to restrict content made on the basis of their content restriction policies.[40]

- Similarly, the **Ranking Digital Rights Project** has undertaken a ranking of the major Internet companies by reference to their compliance with digital rights indicators. These include the following freedom of expression benchmarks: (i) availability of terms of service; (ii) terms of service, notice and record of changes; (iii) reasons for content restriction; (iv) reasons for account or service restriction; (v) notify users of restriction; (vi) process for responding to third-party requests; (vii) data about government requests; (viii) data about private requests; (ix) data about terms of service enforcement; (x) network management (telecommunication companies); and (xi) identity policy (Internet companies).[41]

- Finally, the **Dynamic Coalition on Platform Responsibility** is currently seeking to develop standard Terms and Conditions in line with international human rights standards.[42]

Content-specific principles

Additionally, the special mandates on freedom of expression have issued a number of joint declarations highlighting the responsibilities of States and companies in relation specific content.

- The 2016 **Joint Declaration on Freedom of Expression and Countering Violent Extremism** recommends that States should not subject Internet intermediaries to mandatory orders to remove or otherwise restrict content, except where the content is lawfully restricted in accordance with international standards.[43] Moreover, it is recommended that any initiatives undertaken by

---

were further echoed in the Committee of Ministers' Guide to human rights for Internet users, Recommendation CM/Rec(2014)6 and explanatory memorandum, which states "your Internet service provider and your provider of online content and services have corporate responsibilities to respect your human rights and provide mechanisms to respond to your claims. You should be aware, however, that online service providers, such as social networks, may restrict certain types of content and behaviour due to their content policies. You should be informed of possible restrictions so that you are able to take an informed decision as to whether to use the service or not. This includes specific information on what the online service provider considers as illegal or inappropriate content and behaviour when using the service and how it is dealt with by the provider" (p. 4).

[37] Recommendation CM/Rec (2018) 2 of the Committee of Ministers to member states on the roles and responsibilities of internet intermediaries, adopted by the Committee of Ministers on 7 March 2018 at the 1309th meeting of the Ministers' Deputies

[38] The Manila Principles on Intermediary Liability, March 2015. The Principles have been endorsed by over 50 organisations and over 100 individual signatories

[39] *Ibid.,* Principle IV

[40] *Ibid.*, Principle V c)

[41] Ranking Digital Rights, Corporate Accountability Index, 2015 Research Indicators

[42] Dynamic Coalition on Platform Responsibility is a multi-stakeholder group fostering a cooperative analysis of online platforms' responsibility to respect human rights, while putting forward solutions to protect platform-users' rights.

[43] Joint Declaration on Freedom of Expression and countering violent extremism, adopted by the UN Special

private companies in relation to countering violent extremism should be robustly transparent, so that individuals can reasonably foresee whether content they generate or transmit is likely to be edited, removed or otherwise affected, and whether their user data is likely to be collected, retained or passed to law enforcement authorities.[44]

- The 2017 **Joint declaration on freedom of expression and 'fake news', disinformation and propaganda** recommended, *inter alia,* that intermediaries adopt clear, pre-determined policies governing actions that restrict third-party content (such as deletion or moderation) which go beyond legal requirements.[45] These policies should be based on objectively justifiable criteria rather than ideological or political goals and should, where possible, be adopted after consultation with their users.[46] Intermediaries should also take effective measures to ensure that their users can easily access and understand their policies and practices (including terms of service), and detailed information about how such policies and practices are enforced, and, where relevant, by making available clear, concise and easy to understand summaries of, or explanatory guides to, those policies and practices.[47] It also recommended that intermediaries should respect minimum due process guarantees including by notifying users promptly when content which they create, upload or host may be subject to a content action and by giving the user an opportunity to contest that action.[48]

- The Special Rapporteur on FOE and the Special Rapporteur on violence against women have urged States and companies to address **online gender-based abuse**, whilst warning against censorship.[49] The Special Rapporteur on FOE has highlighted that vaguely formulated laws and regulations that prohibit nudity or obscenity could have a significant and chilling effect on critical discussions about sexuality, gender and reproductive health. Equally, discriminatory enforcement of terms of service on social media and other platforms may disproportionately affect women and those who experience multiple and intersecting discrimination.[50] The special mandate holders recommended that human rights-based responses which could be implemented by governments and others could include education, preventative measures, and steps to tackle the abuse-enabling environments often faced by women online.

The protection of the right to privacy and anonymity online

Guaranteeing the right to privacy in online communications is essential for ensuring that individuals have the confidence to freely exercise their right to freedom of expression.[51]

Rapporteur on FOE, the OSCE Representative on Freedom of the Media, the OAS Special Rapporteur on Freedom of Expression and the ACHPR Special Rapporteur on Freedom of Expression and Access to Information, 4 May 2016, para 2 e)

[44] *Ibid.*, para 2 i)

[45] Joint declaration on freedom of expression and "fake news", disinformation and propaganda, adopted by the UN Special Rapporteur on FOE, the OSCE Representative on Freedom of the Media, the OAS Special Rapporteur on Freedom of Expression and the ACHPR Special Rapporteur on Freedom of Expression and Access to Information, 3 March 2017, para 4 a)

[46] *Ibid.*

[47] *Ibid.*, para 4 b)

[48] *Ibid.*, para 4 c)

[49] Joint Press Release of the UN Special Rapporteurs on FOE and violence against women, UN experts urge States and companies to address online gender-based abuse but warn against censorship, 08 March 2017

[50] Ibid.

[51] The right of private communications is protected in international law through Article 17 of the ICCPR, which provides, *inter alia*, that: "[n]o one shall be subjected to arbitrary or unlawful interference with his privacy, family, home or correspondence, nor to unlawful attacks on his honour and reputation." The UN Special Rapporteur on promotion and protection of human rights and fundamental freedoms while countering terrorism has argued that like restrictions on the right to freedom of expression under Article 19, restrictions of the right to privacy under Article 17 of the ICCPR should be interpreted as subject to the three-part test; see the Report of the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism Martin Scheinin, A/HRC/13/37, 28 December 2009.

The inability to communicate privately substantially affects individuals' freedom of expression rights. In his report of May 2011, the Special Rapporteur on FOE expressed his concerns over the fact that States and private actors use the Internet to monitor and collect information about individuals' communications and activities on the Internet, and that these practices can constitute a violation of Internet users' right to privacy, and ultimately impede the free flow of information and ideas online.[52]

The Special Rapporteur on FOE also recommended that States should ensure that individuals can express themselves anonymously online and refrain from adopting real-name registration systems.[53]

Further, in his May 2015 report on encryption and anonymity in the digital age, the Special Rapporteur on FOE recommended that States refrain from making the identification of users a pre-condition for access to digital communications and online services, and refrain from requiring SIM card registration for mobile users.[54] He also recommended that corporate actors reconsider their own policies that restrict encryption and anonymity (including through the use of pseudonyms).[55]

[52] The May 2011 Report of the Special Rapporteur on FOE, *op.cit.*, para 53
[53] *Ibid.*, para 84
[54] Report of the Special Rapporteur FOE, A/HRC/29/32, 22 May 2015, para 60
[55] *Ibid.*

Analysis of the Twitter Rules

General comments

At the outset, ARTICLE 19 notes that the Twitter Rules, policies and guidelines suffer from two significant shortcomings:

- **Confusing structure**: From ARTICLE 19's perspective, the way in which the Twitter Rules, policies and guidelines are organised on the Twitter website can be somewhat confusing and hard to follow. In particular, the "Twitter Rules and policies" are supplemented by a number of "General guidelines and policies". However, it is unclear how these two categories are different. The "violent threats and glorification of violence guidelines" are relevant to the "hateful conduct policy," however this is not obvious from looking at the "Rules and policies" page.[56]

  In our view, it would be helpful for the Twitter Rules, policies and guidelines to be rationalised and organised in such as a way as to make them more easily accessible to users. In particular, it would be helpful for the rules regarding each type of content to be consolidated, so that Twitter's policy on e.g. 'hate speech' could be found in one place rather than on several separate webpages. Facebook's community standards offer a useful blueprint in this respect, although they could also be further improved.[57] We also note that any re-organisation of the rules would require greater conceptual clarity, particularly as regards the distinctions between 'hate speech,' 'credible threats of violence,' 'harassment' and other types of abuse, for example.

- **Lack of clarity regarding Twitter Rules updates and latest version**: ARTICLE 19 notes that it is very difficult to determine what the latest version of the Twitter Rules, policies and guidelines is. Whilst parts of the website suggest that the latest version was published in November 2017, it appears that further changes have been made since then but this is never made clear.[58] Changing rules with little to no notice is bound to confuse users and creates uncertainty about the standards of conduct applied by Twitter. In our view, the company should clearly identify changes made to its Terms of Service and when the latest updates were made. It might also be useful to consider making the Twitter Rules, policies and guidelines available in one consolidated document with the latest version of these rules for reference purposes.

Recommendations:

- The Twitter Rules, policies and guidelines should be re-organised and consolidated so that the company's rules in relation to particular types of content can be easily found in one place. Consideration should be given also to making the Twitter Rules, policies and guidelines available in one consolidated document;
- Twitter should make clear when the Twitter Rules, policies and guidelines were last updated and identify which parts were amended.

'Hate speech'

Twitter does not use the term 'hate speech.' Instead it prohibits "hateful conduct" in several places:

- In the Twitter Rules which include a section on "abusive behaviour" and sub-section "abuse and hateful conduct;"
- In the Twitter Rules which include a section on "abusive profile information;"

[56] Twitter, General guidelines and policies
[57] See, for example, ARTICLE 19, Facebook Community Standards, July 2018
[58] For instance, we have noticed changes in the Twitter definition of "hateful conduct" between June and August versions of the Twitter Rules; *c.f.* ARTICLE 19, Sidestepping Rights: Regulating Speech by Contract, June 2018.

- In the Twitter Rules Media Policy which includes a section on "hateful imagery;" and
- In a dedicated "Hateful conduct policy" page.

We note some discrepancies and similarities between these materials.

- The Twitter Rules prohibit *inter alia,*
  - **"Hateful conduct**" which is defined as promotion of "violence against, threats, or harassment of other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease;"[59]

  - "**Abusive profile information**" which prohibits the use of username, display name, or profile bio "to engage in abusive behaviour, such as targeted harassment or expressing hate towards a person, group, or protected category." Examples of such behaviour are listed as i) violent threats; ii) abusive slurs; iii) epithets, racist, or sexist tropes; iv) abusive content that reduces someone to less than human; and v) content that incites fear;[60]

  - **The use of hateful images or symbols** in users' profile image or profile header.[61] Under the Rules, "hateful imagery" are "logos, symbols, or images whose purpose is to promote hostility and malice against others based on their race, religion, disability, sexual orientation, or ethnicity/national origin." Examples include i) symbols historically associated with hate groups (for example, the Nazi swastika); ii) images depicting others as less than human or altered to include hateful symbols; and iii) altered image references to a mass murder that targeted a protected category.

- The Hateful Conduct policy:
  - Prohibits "[promoting] violence against or directly [attacking] or [threatening] other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease;"[62]

  - States that Twitter does "not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories;"[63]

  - Lists examples of "hateful conduct" – these include "violent threats; wishes for the physical harm, death, or disease of individuals or groups; references to mass murder, violent events, or specific means of violence in which/with which such groups have been the primary targets or victims; behaviour that incites fear about a protected group; repeated and/or or non-consensual slurs, epithets, racist and sexist tropes, or other content that degrades someone."[64]

In assessing "abusive behaviour" and determining appropriate enforcement actions, the Twitter Rules make clear that Twitter takes context into account and in particular the following factors: [65]

- the behaviour is targeted at an individual or group of people;
- the report has been filed by the target of the abuse or a bystander;
- the behaviour is newsworthy and in the legitimate public interest.

[59] Twitter Rules, op.cit., Abusive Behaviour section. Prohibited abusive behaviour includes: (i) violence and physical harm; (ii) abuse and hateful conduct, (iii) private information and intimate media; and (iv) impersonation."
[60] Twitter Rules, Abusive Profile Information
[61] Twitter Rules, Media Policy
[62] Twitter, Hateful Conduct Policy
[63] *Ibid.*
[64] *Ibid.*
[65] Additional factors are mentioned in the Twitter General guidelines and policies, Our enforcement philosophy

ARTICLE 19 welcomes Twitter's attempt at differentiating different types of "abusive behaviour", including the distinction between merely "abusive" and "hateful" conduct. In the absence of more information on how these policies are enforced and how the severity of the conduct at issue is assessed in practice, however, ARTICLE 19 notes that they allow for broad interpretation and restrictions beyond international standards on freedom of expression.

Although international law does not define 'hate speech' *per se,* the expression of hatred towards an individual or group on the basis of a protected characteristic can be divided into three categories, distinguished by the response international human rights law requires from States:[66]

- Severe forms of 'hate speech' that international law *requires* States to prohibit, including through criminal, civil, and administrative measures, under both international criminal law and Article 20(2) of the ICCPR;
- Other forms of 'hate speech' that States *may* prohibit to protect the rights of others under Article 19(3) of the ICCPR, such as discriminatory or bias-motivated threats or harassment;
- Lawful 'hate speech' that should be permitted but nevertheless raises concerns in terms of intolerance and discrimination and therefore deserves a critical response by the State.

While Twitter, as a company, would not be expected to adopt the same types of measures as States, the above categories should guide its response to 'hateful conduct.' In our view, it would also help clarify the various concepts that come into play. For instance, Twitter uses a number of ill-defined terms such as "hateful conduct," "hate," "abuse" and overlapping concepts, including "hateful conduct," "harassment" and "threats." The distinction between "hateful conduct" and "hateful imagery and display names" is not entirely clear. In any event, it conflates harassment with incitement to violence and/or discrimination.

Although it is positive that Twitter highlights the importance of context in making decisions about the enforcement of its policies, it should provide a more detailed list of the various factors that might come into play. For instance, there is no reference to parody or humour as exemptions despite the fact that these are important considerations when considering complaints.[67] More generally, it is regrettable that Twitter does not provide any examples of how the policies are implemented in practice as it would go a long way towards explaining how the company takes context into consideration.

Recommendations:

- Twitter's policies of "hateful conduct" should be clearly presented and should be more closely aligned with international standards on freedom of expression, including by differentiating between different types of prohibited expression on the basis of severity;
- Twitter should provide case studies or more detailed examples of the way in which it applies its "hateful conduct" policies.

Extremism/Terrorism

Twitter does not have a specific policy to deal with 'terrorist' content; rather, several policies are relevant to this type of content.

- First, the Twitter Rules prohibit the use of Twitter "for any unlawful purposes or in furtherance of illegal activities."[68] Moreover, users "may not make specific threats of violence or wish for the serious physical harm, death, or disease of an individual or group of people" which includes,

[66] *C.f.* ARTICLE 19, Hate Speech Explained: A Toolkit, 2015
[67] Though parody is the subject of a dedicated, see Twitter, Parody, newsfeed, commentary, and fan account policy
[68] Twitter Rules, *op.cit.*

inter alia, "threatening or promoting terrorism."[69] Users are also prohibited from affiliating themselves with organisations that – whether by their own statements or activities on or off the platform – "use or promote violence against civilians to further their causes."[70]

- Second, the section on "violent threats and glorification of violence" in the Twitter General guidelines and policies further explains that Twitter "will not tolerate behaviour that encourages or incites violence against a specific person or group of people" and that it will "also take action against content that glorifies acts of violence in a manner that may inspire others to replicate those violent acts and cause real offline danger, or where people were targeted because of their potential membership in a protected category."[71] Twitter understands the "glorification of violence" as "behaviour that condones or celebrates violence and/or its perpetrators in a manner that may promote imitation of the act." It also prohibits the glorification of violence where protected categories have been the primary target or victim. The "glorification" of terrorist attacks, rape, sexual assault and mass murders are given as examples of violations of Twitter's policy.

- Thirdly, the section on "violent extremist groups" in the Twitter General guidelines and policies makes clear that such groups are prohibited from using Twitter's services.[72] Twitter defines "violent extremist groups" as groups that: (i) identify through their stated purpose, publications, or actions, as an extremist group; (ii) have engaged in, or currently engage in, violence (and/or the promotion of violence) as a means to further their cause; (iii) target civilians in their acts (and/or promotion) of violence.[73] Twitter goes on to explain that exceptions will be considered for groups that have reformed or are currently engaged in a peaceful resolution process, as well as groups with representatives elected to public office through democratic elections. The policy does not apply to military or government entities.

- In addition, Twitter explains that users will be deemed affiliates of terrorist groups if they: (i) state or suggest that an account represents or is part of a violent extremist group; (ii) provide or distribute services (e.g., financial, media/propaganda) in furtherance of progressing a violent extremist group's stated goals; (iii) engage in or promoting acts for the violent extremist group; and (iv) recruit for the violent extremist group.[74]

ARTICLE 19 notes that although Twitter attempts to explain its understanding of several relevant terms, its definitions are inconsistent with international standards on freedom of expression:

- Rather than relying on the international definition of incitement to terrorism, Twitter uses vague and overbroad language such as "glorification" or "violent extremism," which are inconsistent with international standards on freedom of expression. For example, the international mandates on freedom of expression and counter-terrorism have highlighted that the prohibition on incitement to terrorism should avoid references to vague terms such as the 'promotion' or 'glorification' of terrorism.[75]

- Secondly, Twitter's definitions and policies on violent extremism rely on a lower threshold of likelihood of violence occurring than under international law. In particular, Twitter prohibits behaviour that "condones or celebrates violence and/or its perpetrators in a manner that *may* promote imitation of the act" (our emphasis). Under international law, however, there should be an objective risk that the act incited will be committed and intent that the message at

[69] *Ibid*, Violence and physical harm sub-section
[70] *Ibid.*
[71] Twitter, Violent threats and glorification of violence
[72] Twitter, Violent extremist groups
[73] *Ibid.*
[74] *Ibid.*
[75] See e.g. Report of the Special Rapporteur on FOE, August 2011, A/66/290, para. 34

issue incites the commission of a terrorist act for content to amount to incitement to terrorism.[76]

We further note that it is understandable that Twitter wishes to deny its platform to "violent extremist groups" and that this is not necessarily an unreasonable restriction on freedom of expression in and of itself. However, a key difficulty is the lack of agreed definition of terrorism at international level. This is compounded by Twitter's lack of a more comprehensive definition of terrorist activity beyond the use of "violence" and the targeting of civilians. For instance, the UN Special Rapporteur on Counter-Terrorism suggested that any definition of terrorism should include a reference to the 'intentional taking of hostages,' 'actions intended to cause death or serious bodily injury to one or more members of the general population or segments of it' or 'actions involving lethal or serious violence to one or more members of the general population or segments of it." [77]

Another difficulty is the lack of clarity around the way in which Twitter deals with individuals designated as 'terrorist' by certain governments but who may otherwise be regarded as freedom fighters. For instance, Twitter does not explicitly say whether it complies with the US State Department list of designated terrorist groups. If that is the case, this can be a problem as this list includes groups, which are not designated as 'terrorist' by the UN, such as the Kurdistan Workers' Party (PKK). It is also unclear how Twitter deals with 'terrorist' lists compiled by other governments. Again, this is a problem as a government's terrorist group may well be regarded as a social movement or (e.g. indigenous) group with legitimate claims.

While Twitter's policy on violent extremist groups contains welcome exceptions in relation to groups that have "reformed" or are engaged in a peaceful resolution process, it remains unclear how this is articulated with the various, potentially conflicting, legal requirements outlined above. In any event, Twitter's exceptions appear to be insufficient to deal with individuals and groups that may be considered as freedom fighters, or social movements. For instance, Nelson Mandela and the African National Congress would arguably have been barred from the platform during the Apartheid regime.

More generally, as noted earlier, we regret that Twitter does not give concrete examples of how its standards are applied in practice. This would help clarify some of the concerns outlined above.

Recommendations:

- Twitter should align its definition of terrorism and incitement to terrorism with that recommended by the UN Special Rapporteur on counter-terrorism. In particular, it should avoid the use of vague terms such as "violent extremism," "celebrate," "condone," "glorification" or "promotion" of terrorism;
- Twitter should give examples of organisations falling within the definition of "violent extremist groups." In particular, it should explain how it complies with various governments' designated lists of terrorist organisations, particularly in circumstances where certain groups designated as 'terrorist' by one government may be considered as legitimate (e.g. freedom fighters) by others.
- Twitter should provide case studies explaining how it applies its standards in practice.

Privacy and morality-based restrictions

Twitter, like most major social media companies, prohibits various types of content that would constitute a restriction on freedom of expression on the grounds of public morals or the protection of the right to privacy. This content generally falls within three categories: threats of violence/harassment; adult content; and the posting of private information.

[76] See e.g. the model offence of incitement to terrorism in the UN Special Rapporteur on counter-terrorism's report, 22 December 2010, A/HRC/16/51, paras 29-32
[77] *Ibid.*

Whilst these categories encompass material that is clearly unlawful (such as credible threats of physical violence or harassment), they can also include lawful content (such as pornography, or offensive or insulting content that falls short of harassment). Other types of content may fall in a grey area, when their publication constitutes an interference with the right to privacy but may be otherwise justified in the public interest.

Violence and physical harm, 'abuse' and 'unwanted sexual advances'

The Twitter Rules protects the right to privacy of its users under a number of headings, including "violence and physical harm," "abuse" and "unwanted sexual advances:"

- The Twitter Rules on "violence and physical harm" provide that users may not issue "specific threats of violence or wish for the serious physical harm, death, or disease of an individual or group of people." Furthermore, users "may not promote or encourage suicide or self-harm." [78]

- The Twitter Rules on "abuse and hateful conduct" define "abuse" as the targeted harassment of someone, or inciting other people to do so. Twitter considers abusive behaviour any "attempt to harass, intimidate, or silence someone else's voice."[79]

- Finally, the Twitter Rules on "abuse and hateful conduct" also prohibit "unwanted sexual advances" defined as "directing abuse at someone by sending unwanted sexual content, objectifying them in a sexually explicit manner, or otherwise engaging in sexual misconduct."[80]

ARTICLE 19 notes that the Twitter Rules on violence and physical harm tend to be drafted in relatively broad language since they include mere "wishes" for the serious physical harm, death, or disease of an individual, even though those may not be meant seriously. Equally, the Rules do not specify that threats of violence must be credible before Twitter will take enforcement action. More generally, the Rules do not refer to a requirement of intent in relation to incitement to violence, which is inconsistent with international standards on freedom of expression.

ARTICLE 19 further notes that Twitter's description of "harassment" as a form of "abuse" is somewhat confusing. "Abuse" is a very broad term that is generally understood to cover behaviour falling below the threshold of criminality. By contrast, harassment generally entails conduct causing "alarm or distress." However, these elements are not mentioned in the definition. It would also have been helpful for Twitter to explain how offensive content might be distinguished from harassment or other "hateful conduct" or "abuse."

More generally, ARTICLE 19 believes that Twitter should highlight in its Rules that offensive content will not automatically be taken down unless it violates other rules. Whilst it is open to companies to take down purely abusive or even offensive content, this should not be at the detriment of public debate, particularly on matters of public interest.

Graphic violence and adult content

Twitter allows some forms of graphic violence and/or adult content on its platforms subject to some restrictions.

Graphic violence is defined as "any form of gory media related to death, serious injury, violence, or surgical procedures."[81] The media policy lists depictions of moments at which someone dies,

[78] The Twitter Rules, *op.cit.*
[79] The Twitter Rules, *op.cit.*
[80] The Twitter Rules, *op.cit.*
[81] Twitter, Enforcing our rules, Sensitive content; and Twitter, Media Policy, *op.cit.*

gruesome crime or accident scenes, bodily harm, torture, dismemberment or mutilation as examples. Twitter further defines adult content as "any media that is pornographic and/or may be intended to cause sexual arousal." The media policy goes on to give a non-exhaustive list of examples, including depictions of:

- full or partial nudity (including close-ups of genitals, buttocks, or breasts).
- simulating a sexual act
- intercourse or any sexual act (may involve humans, humanoid animals, cartoons, or anime)

The above content may be published on Twitter; however, it may only be published in Tweets marked as containing sensitive media. Moreover, users are not allowed to use such content in their profile or header images. Additionally, Twitter may sometimes require users to remove excessively graphic violence out of respect for the deceased and their families if it receives a request from their family or an authorised representative.

At the same time, Twitter allows for some exceptions. In relation to full or partial nudity, exceptions may be made for artistic, medical, health, or educational content. It further specifies that breastfeeding content does not need to be marked as sensitive.

ARTICLE 19 notes that Twitter's policy on adult and graphic content appears to be relatively permissive, particularly compared to other social media platforms (e.g. Facebook's policy on nudity). ARTICLE 19 notes however that it would be helpful for Twitter to provide examples of how the policy is applied in practice. For instance, it is unclear how its policy on graphic violence may be applied in the context of news reporting. Moreover, it is worth noting that individuals who are not sufficiently tech-savvy to amend the settings of their account to see sensitive media may be denied an opportunity to gain access to information, which may otherwise be in the public interest.

Other personal or private information

The "private information and intimate media" section of the Twitter Rules is divided into three sections:

- **Private information**: under this heading, the Twitter Rules provide that "users may not publish or post other people's private information without their express authorization and permission. Definitions of private information may vary depending on local laws." The Twitter General guidelines and policies go on to specify that private information for the purposes of the Twitter Rules include, but are not limited to, credit card information, social security or other national identity numbers, private residences, personal home addresses, or other locations that are considered private, non-public personal phone numbers and non-public personal email addresses.[82] It does not specifically mention personal medical records.

- **Intimate media**: under this heading, users are prohibited from posting or sharing intimate photos or videos of someone that were produced or distributed without their consent. The Twitter intimate media policy goes on to give a non-exhaustive list of examples of intimate media that violate the policy:[83]
  - hidden camera content involving nudity, partial nudity, and/or sexual acts;
  - images or videos that appear to have been taken secretly and in a way that allows the user to see the other person's genitals, buttocks, or breasts (content sometimes referred to "creepshots" or "upskirts");
  - images or videos captured in a private setting and not intended for public distribution;
  - images or videos that are considered and treated as private under applicable laws.

- **Threats to expose/hack:** The Twitter Rules prohibit users from threatening to expose someone's

[82] Twitter, About Private Information on Twitter
[83] Twitter, About Intimate Media on Twitter

private information or intimate media. Users also may not threaten to hack or break into someone's digital information.

ARTICLE 19 notes that the Twitter Rules on private information are generally consistent with the protection of the right to privacy. However, they fail to specify that these rules may need to be balanced with the protection of the right to freedom of expression in relation to information that may be in the public interest.

Recommendations:

- Twitter should explain in more detail the relationship between threats, harassment, and online abuse and distinguish this from offensive content (which should not be limited as such);
- Twitter should provide detailed examples or case studies of the way in which it applies its standards in practice, including with a view to ensuring protections for minority and vulnerable groups;
- Twitter should state more clearly that offensive content will not be taken down as a matter of principle unless it violates other rules;
- Twitter should make more explicit reference to the need to balance the protection of the right to privacy with the right to freedom of expression. In so doing, it should also refer to the criteria developed, *inter alia,* in the Global Principles on the Protection of Freedom of Expression and Privacy.

'Fake news'

Twitter does not explicitly ban 'fake news' or "false information" on its platform but a number of its policies, on impersonation,[84] spam,[85] and bots[86] may be applicable.

Whilst Twitter has stated that it does not wish to be an arbiter of truth, it has recently stepped up its crackdown on some Russian "fake" accounts that allegedly interfered in the US election.[87] These efforts rely on the company's internal "systems" to detect "suspicious" activity on the platform, including suspicious accounts, tweets, logins and engagement. The company also states that it is investing in systems to stop "bad content" at its source.[88] However, the company does not disclose how these "systems" are used in order to prevent "bad actors" from gaming the system, nor does it explain what criteria are used to determine who is a "bad actor."[89] Most recently, Twitter said that it was increasingly using automated and proactive detection methods to find "misuses" of its platform before they impact its users' experience. However, the company does not explain what constitutes a "misuse" or "abuse" of its platform.[90]

In ARTICLE 19's view, Twitter's policies on spam, bots and fake accounts are broadly consistent with international standards in this area. However, the company relies on a number of very loose terms such as "suspicious" or "bad actors" or "malicious behaviour" so that it is unclear whether - and if so, to what extent - the company removes "false information" in practice. Moreover, very little information is available about closed accounts on grounds of "suspicious" activity.

Recommendations:

- Twitter should be more transparent and explain in more detail how its algorithms detect 'fake' accounts, including by listing the criteria on the basis of which these algorithms operate.

[84] Twitter, Impersonation Policy
[85] Twitter, About specific instances when a Tweet's reach may be limited
[86] Twitter, Our Approach to Bots & Misinformation, 14 June 2017
[87] Twitter Public Policy, Update: Russian Interference in 2016 US Election, Bots, & Misinformation, 28 September 2017
[88] *Ibid.*
[89] *Ibid.*
[90] See Twitter, How Twitter is fighting spam and malicious automation, 26 June 2018.

- Twitter should explain how the measures it is adopting to fight fake accounts, bots etc. are different from removing false information.
- The company should also define of further define what it considers to be "suspicious activity," "bad actors" or "platform manipulation."

Content removal processes

Like other major social media companies, Twitter relies on filters or "automated detection methods" to remove certain types of content on its own initiative. It also provides for different types of reporting mechanisms, depending on the nature of the complaint at issue.[91] Reporting forms are generally comprehensive.[92] It is unclear, however, whether Twitter relies on a trusted flagger system.[93]

Similarly, it is difficult to find information about any appeals mechanism to challenge a decision taken by Twitter in relation to either an account or particular piece of content. Rather than providing a separate page dealing with such mechanisms, a link is provided on a seemingly *ad hoc* basis at the end of some of its content-related policies.[94] Therefore, it appears that any individual whose content is removed on the basis of Twitter's policies is generally not given any reasons for the decision, or a clear opportunity to appeal. Appeals only seem to be available when accounts are suspended.[95]

Positively, it appears that Twitter generally makes good faith efforts to inform users about legal requests it receives to remove content.[96] However, generally speaking, reasons for actions taken by Twitter, or access to an appeals mechanism do not appear to be given on a consistent or systematic basis. In our view, these are significant shortfalls in Twitter's internal processes and inconsistent with due process safeguards.

Recommendations:

- Twitter should ensure that its appeals process complies with the Manila Principles on Intermediary Liability, particularly as regards notice, the giving of reasons, and appeals processes;
- Twitter should be more transparent about its use of algorithms to detect various types of content, such as 'terrorist' videos, 'fake' accounts or 'hate speech;'
- Twitter should clarify whether it relies on a trusted flagger system, and if so, it should provide information about it including by identifying members of the scheme and the criteria being applied to join it.

Sanctions

Twitter explains in its Twitter Rules that "all individuals accessing or using Twitter's services must adhere to the policies set forth in the Twitter Rules. Failure to do so may result in Twitter taking one or more of the following enforcement actions: (i) requiring you to delete prohibited content before you can again create new posts and interact with other Twitter users; (ii) temporarily limiting your ability to create posts or interact with other Twitter users; (iii) asking you to verify account ownership with a phone number or email address; or (iv) permanently suspending your account(s).[97] Twitter also has a dedicated page explaining its enforcement options at various levels, from response to

[91] Twitter Help Centre, Report violations
[92] Twitter Help Centre, Report abusive behavior
[93] Twitter relies on Trust and Safety Partners who are members of its Trust and Safety Council. However, it is unclear whether the Council operates as a trusted flagger system
[94] Violent extremist groups, *op.cit.* and Twitter Help Centre, Appeal an account suspension or locked account
[95] *Ibid.*
[96] Twitter Help Centre, Content removal requests
[97] Twitter Rules, *op.cit.*

tweets to direct messages and accounts.[98] It also clearly explains the general principles guiding the enforcement of its policies.[99]

In our view, the above Terms are broadly consistent with international standards on freedom of expression and Manila Principles on Intermediary Liability. These standards provide that content restriction policies and practices must comply with the tests of necessity and proportionality under human rights law. At the same time, we regret that Twitter, like many other similar companies, appears to be increasingly applying country filters so that the promise of free expression 'beyond borders' is rapidly evaporating.[100] We also note that very little information is available as to the way in which Twitter applies its sanctions policy in practice.

Recommendations:

- Twitter should provide case studies of the way in which it applies its sanctions policy;
- Twitter should provide disaggregated data on the types of sanctions it applies in its Transparency Report.

[98] Twitter Help Centre, Our range of enforcement options
[99] Twitter Help Centre, Our approach to policy development and enforcement philosophy
[100] See e.g. EFF's report, Who has your back? Censorship Edition, 2018. See also, Twitter, *About Country Withheld Content:* https://help.twitter.com/en/rules-and-policies/tweet-withheld-by-country

About ARTICLE 19

ARTICLE 19 advocates for the development of progressive standards on freedom of expression and freedom of information at international and regional levels, and their implementation in domestic legal systems. The Law Programme has produced a number of standard-setting publications which outline international and comparative law and best practice in areas such as defamation law, freedom of expression and equality, access to information, and broadcast regulation.

On the basis of these publications and ARTICLE 19's overall legal expertise, the organisation publishes a number of legal analyses each year and comments on legislative proposals and existing laws that affect the right to freedom of expression. This analytical work, carried out since 1998 as a means of supporting positive law reform efforts worldwide, frequently leads to substantial improvements in proposed or existing domestic legislation. All of our analyses are available at https://www.article19.org/law-and-policy.

If you would like to discuss this analysis further, or if you have a matter you would like to bring to the attention of the ARTICLE 19 Law and Policy Team, you can contact us by email at legal@article19.org.