ARTICLE 19

Facebook Community Standards

June 2018

Legal analysis

Executive summary

In June 2018, ARTICLE 19 reviewed the compatibility of Facebook's Community Standards with international standards on freedom of expression. Facebook published the latest version of its community standards in April 2018.

Facebook Community Standards are divided into six sections: (i) violence and criminal behaviour; (ii) safety; (iii) objectionable content; (iv) integrity and authenticity; (v) respecting intellectual property; and (vi) content-related requests. Although the latest version of the Community Standards is much more transparent and detailed than previous iterations, our analysis shows that they continue to fall below international standards on freedom of expression. This is especially the case of the Community Standards on 'hate speech,' 'terrorism' and bullying and harassment, as well as Facebook's content removal procedures.

ARTICLE 19 encourages Facebook to bring its Community Standards 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. Facebook's definition of 'hate speech' should be more closely aligned with international standards on freedom of expression, including Article 20(2) ICCPR which requires States to prohibit the advocacy of hatred that incites to discrimination, hostility or violence. Also, Facebook should provide case studies or more detailed examples of the way in which it applies its policies on 'hate speech.' Failure to express intent in relation to 'hate speech' (whether educational or otherwise) should not lead to automatic removal of that content in practice;

2. The definition of 'hate organisations' and 'attack' should be narrowed, including by adding a requirement of intent to harm particular groups;

3. Facebook should align its definition of 'terrorism' with that recommended by the UN Special Rapporteur on counter-terrorism. In particular, it should avoid the use of vague terms such as 'praise,' 'express support,' 'glorification' or 'promotion';

4. Facebook should give examples of organisations falling within its definition of terrorist organisation. 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 (e.g. on beheading videos);

5. Facebook should strive to narrow its definitions of bullying and harassment in order to prevent legitimate content from being removed. It should ensure that the definition of bullying and harassment remain distinct. Facebook should explain in more detail the relationship between "threats," "harassment," and "online abuse"/"bullying" and distinguish these from "offensive content" (which should not be limited as such). Further, Facebook 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;

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

7. Facebook 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 make reference to the criteria developed, *inter alia,* in the Global Principles on the Protection of Freedom of Expression and Privacy;

8. Facebook should improve its voluntary initiatives aimed at tackling 'fake news' by including a charter of ethics comparable to the highest professional standards of journalism and involving a wide range of stakeholders;

9. Facebook should explain in more detail how its algorithms detect 'fake accounts' or produce more 'reliable' results, including by listing the criteria on the basis of which these algorithms operate;

10. Facebook 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;

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

12. Facebook should provide information about its trusted flagger system, including identifying members of the scheme and the criteria being applied to join it;

13. Facebook should refrain from putting in place contact points in countries with a poor record on the protection of freedom of expression;

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

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

16. Facebook should stop requiring its users to use their real-name. It should not require users to prove their identity.

Table of contents

Introduction

In this analysis, ARTICLE 19 reviews the latest version of Facebook's Community Standards, which were published in April 2018.

Since its inception in 2006, Facebook has grown into a multi-billion-pound company, with over two billion users. As such, it has become a critical gateway for the exercise of freedom of expression online. By the same token, it has also become a major player in the regulation and moderation of online content.

In recent years, Facebook has become more transparent about the rules it applies to the removal of content on its platform. Facebook Community Standards are divided into six sections:

- Violence and criminal behaviour;
- Safety;
- Objectionable content;
- Integrity and authenticity;
- Respecting intellectual property; and
- Content-related requests.

ARTICLE 19 finds that the 2018 version of the Community Standards is much more detailed and provides a much better insight into the types of factors taken into account by Facebook's content moderators. Facebook also now publishes a Transparency Report, including about the enforcement of its community standards. These are welcome developments.

However, ARTICLE 19 concludes that Facebook's Community Standards continue to fall below international standards on freedom of expression. In particular, Facebook often imposes broader restrictions on content than would otherwise be found in legislation (e.g. on bullying) or international standards on 'hate speech.' Most rules also remain very broad in scope, leaving significant discretion to Facebook in their implementation. As such, they are highly likely to lead to inconsistent application. This is all the more so given that Facebook does not provide case studies or detailed examples of its internal 'case-law.'

ARTICLE 19 believes that social media companies, including Facebook, should respect international standards on human rights consistent with the UN Guiding Principles on Business & 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 Facebook, which occupy such a prominent position in the Internet ecosystem.

Our analysis is divided into two parts. First, we set out international standards on freedom of expression that companies should respect, consistent with the UN Guiding Principles. Secondly, we analyse Facebook's Community Standards in some key areas, focusing on 'hate speech,' 'terrorist' content, privacy and morality-based restrictions on content, and 'fake news.' We also examine Facebook's content removal processes, sanctions, and real-name and identification policies. Each section contains recommendations on how to bring Facebook's Community Standards in line with international standards on freedom of expression.

International human rights standards

ARTICLE 19's comments on Facebook's Community Standards 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),[1] 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 to 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.[2]

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.[3]

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.[4]

---

[1] Through its adoption in a resolution of the UN General Assembly, the UDHR is not strictly binding on states. However, 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).

[2] See HR Committee, General Comment No. 34 on Article 19: Freedoms of opinion and expression, CCPR/C/GC/34, 12 September 2011, para 11.

[3] HR Committee, *Belichkin v. Belarus*, Communication No. 1022/2001, U.N. Doc. CCPR/C/85/D/1022/2001 (2005).

[4] General Comment No. 34, *op.cit.*, para 43. The General Comment states that "any restrictions on the operation of websites, blogs or any other internet-based, electronic or other such information dissemination system, including systems to support such communication, such as internet service providers or search engines, are only permissible to the extent that they are compatible with paragraph 3. Permissible restrictions generally should be content-specific; generic bans on the operation of certain sites and systems are not compatible with paragraph 3. It is also inconsistent with paragraph 3 to prohibit a site or an information dissemination system from publishing material solely on the basis that it may be critical of the government or the political social system espoused by the government."

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.

Intermediary liability

The four special mandates on freedom of expression have recognised for some time that immunity from liability was 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.[5]

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 its behalf.[6] 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.[7]

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.[8] 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:[9]
  - 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;
  - Track and communicate performance, risks and government demands; and
  - Make remedies available where adverse human rights impacts are created.

[5] The 2011 Joint Declaration, *op. cit.*
[6] The Report of the Special Rapporteur on FoE, 16 May 2011, A/HRC/17/27, para 43.
[7] *Ibid.*, para 42.
[8] 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.
[9] *Ibid.*, Principle 15.

- In his **May 2011 report to the Human Rights Council,** the Special Rapporteur on FoE highlighted that, while States are the duty-bearers for human rights, Internet intermediaries also have a responsibility to respect human rights and referenced the UN Guiding Principles in this regard.[10] 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.[11] 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.[12] 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.[13]

- In his **June 2016 Report to the Human Rights Council**,[14] 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,"[15] 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."[16]

- 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[17] 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.[18] He also encouraged companies to publish transparency reports about government requests for user data or content removal;[19] challenge requests for content removal or requests for user data that may violate the law or internationally recognised human rights;[20] notify individuals affected by any measure restricting their freedom of expression and provide them with non-judicial remedies,[21] and; take proactive protective measures to develop good business practices consistent with respect for human rights.[22]

---

[10] The May 2011 Report of the Special Rapporteur on FoE, *op.cit.*, para 45.
[11] *Ibid.* para 46.
[12] *Ibid.*, paras 47 and 76.
[13] *Ibid.*, paras 48 and 77.
[14] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, 11 May 2016, A/HRC/32/38; para 40 – 44.
[15] *Ibid.*
[16] *Ibid.*, para 43.
[17] OAS Special Rapporteur on FoE, Freedom of Expression and the Internet, 2013. The report noted 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;" paras 110-116.
[18] *Ibid.*, paras 111-112.
[19] *Ibid.*, para 113.
[20] *Ibid.*, para 114.
[21] *Ibid.*, para 115.
[22] *Ibid.*, para 116.

- In the **2016 report on Standards for a Free, Open and Inclusive Internet**,[23] the OAS Special Rapporteur on FoE recommended that, *inter alia,* companies make a formal and high-level commitment to respect human rights, and back this commitment up 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.[24] He also highlighted the need for companies' policies, operating procedures and practices to be transparent.[25]

- At 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 should stop relying on private companies that control the Internet to impose restrictions that violate States' human rights obligations.[26] He recommended that further guidance should be developed on the responsibilities of business enterprises 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.[27]

- Similarly the Committee of Ministers of the Council of Europe, in its **Recommendation on the protection of human rights with regard to social networking services**, recommended that social media companies should respect human rights and the rule of law, including procedural safeguards.[28] Moreover, in its March 2018 **Recommendation on the roles and 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.[29] 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.[30] In particular, they make clear that companies' content restriction practices must comply with the tests of necessity and

[23] OAS Special Rapporteur on FoE, Standards for a Free, Open and Inclusive Internet, 2016, paras 95-101.
[24] *Ibid.*, para 98.
[25] *Ibid.*, para 99.
[26] 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.
[27] *Ibid.*, p. 24.
[28] 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 were further echoed in the Committee of Ministers Guide to Human Rights for Internet users, 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" Guide to human rights for Internet users, Recommendation CM/Rec(2014)6 and explanatory memorandum, p. 4.
[29] 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.
[30] The Manila Principles on Intermediary Liability, March 2015. The Principles have been endorsed by over 50 organisations and over 100 individual signatories.

proportionality under human rights law,[31] and that intermediaries should provide users with complaints mechanisms to review decisions to restrict content made on the basis of their content restriction policies.[32]

- 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); (xi) identity policy (Internet companies).[33]

- Finally, the **Dynamic Coalition on Platform Responsibility** is currently seeking to develop standard Terms and Conditions in line with international human rights standards.[34]

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.[35] Moreover, they recommended that any initiatives undertaken by 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.[36]

- The 2017 **Joint Declaration on '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 goes beyond legal requirements.[37] 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.[38] 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.[39] It also recommended that intermediaries should respect

[31] *Ibid.,* Principle IV.
[32] *Ibid.*, Principle V c).
[33] Ranking Digital Rights, Corporate Accountability Index, 2015 Research Indicators.
[34] 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.
[35] Joint Declaration on Freedom of Expression and countering violent extremism, 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, 4 May 2016, para 2 e).
[36] *Ibid.*, para 2 i).
[37] The 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 FoE and the ACHPR Special Rapporteur on Freedom of Expression and Access to Information, 3 March 2017, para 4 a).
[38] *Ibid.*
[39] *Ibid.*, para 4 b).

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.[40]

- 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.[41] 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.[42] 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.[43]

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.[44]

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.[45]

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 from requiring SIM card registration for mobile users.[46] He also recommended that corporate actors reconsider their own policies that restrict encryption and anonymity (including through the use of pseudonyms).[47]

---

[40] *Ibid.*, para 4 c).
[41] The 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.
[42] *Ibid.*
[43] The right of private communications is protected in international law through Article 17 of the ICCPR, *op.cit.*, which provides, *inter alia*, that: "No 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.
[44] The May 2011 Report of the Special Rapporteur on FoE, *op.cit.*, para 53.
[45] *Ibid.*, para 84.
[46] Report of the Special Rapporteur to the Human Rights Council on the use of encryption and anonymity to exercise the rights to freedom of opinion and expression in the digital age, A/HRC/29/32, 22 May 2015, para 60.
[47] *Ibid.*

Analysis of the Facebook Community Standards

'Hate speech'

Facebook updated its Community Standards on 'hate speech' in 2015[48] and again in 2018 in an attempt to clarify the criteria that the company uses to remove content.[49] ARTICLE 19 notes that the 2018 version is far more detailed than previous iterations; in particular, it sets out more criteria than before. However, Facebook's approach to 'hate speech' goes beyond the criteria laid down under international law.

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:[50]

- 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 Facebook, as a company, would not be expected to adopt the same types of measures as States, the above categories should guide its response to 'hate speech.' Unfortunately, that is not the case for the following reasons.

Overbroad definition of 'hate speech'

The 'hate speech' section of the Facebook Community Standards, which falls under the sub-heading 'objectionable content', explains that Facebook does not allow 'hate speech.' 'Hate speech' is defined as:

> "A direct attack on people based on what we call protected characteristics – race, ethnicity, national origin, religious affiliation, sexual orientation, sex, gender, gender identity and serious disability or disease. We also provide some protections for immigration status."

While the term 'attack' itself is excessively vague, Facebook defines it as "violent or dehumanising speech, statements of inferiority, or calls for exclusion or segregation."[51]

This is again incredibly broad and would include perfectly lawful statements. However, Facebook further explains these terms as follows:

- Violent speech includes support for death/disease/harm;

[48] Op. cit.
[49] Facebook standards on hate speech as of April 2018 are available from here.
[50] For a full explanation of ARTICLE 19's policy on 'hate speech', see Hate Speech Explained: A Toolkit, 2015.
[51] Facebook Community Standards, Objectionable Content.

- Dehumanising speech includes reference or comparison with filth, bacteria, disease or faeces, reference or comparison to sub-humanity, reference or comparison with animals that are culturally perceived as physically or intellectually inferior;

- Statements of inferiority include a statement or term implying a person's or a group's physical (e.g. "ugly" "hideous" "deformed"), mental ("retarded," "stupid," "idiot") or moral ("slutty," "cheap," "free-riders") deficiency, statements of contempt ("I hate," "I don't like," "X are the worst"); expressions of disgust ("vile," "gross," "disgusting") or cursing at people who share a protected characteristic.

- Users are also not allowed to post content that describes or negatively targets people with slurs, where slurs are defined as words that are commonly used as insulting labels for the above-listed characteristics.

In ARTICLE 19's view, these categories remain overly broad. As a result, legitimate, albeit offensive, expression is highly likely to be removed.

Failure to differentiate between types of 'hate speech'

Facebook further fails to differentiate between different types of 'hate speech.' For instance, 'attack' is broadly defined as encompassing "violent speech," "dehumanising statements" or "statements of inferiority" without making any reference to either the *intent* of the speaker to incite others to take *action*, or the likelihood of a specific type of harm occurring as a result of the speech at issue. The examples given suggest that many different types of legitimate speech are likely to be removed (e.g. statements of contempt or insults that fall short of calling for violence or discrimination). Although Facebook understandably seeks to create a 'safe' environment for its users, it effectively sets a very low bar for free expression, where offensive views are likely to be removed. Moreover, the lack of sufficiently detailed examples (e.g. case studies) means that it is highly likely that Facebook's application of its policies will continue to appear arbitrary and biased.

Dangerous individual and organisations

In addition to the 'hate speech' section, Facebook Community Standards contain a separate section on "Dangerous individuals and organisations." In this section, Facebook states that it does not allow any organisations or individuals that are engaged in organised hate.[52] 'Hate organisations' are defined as:

> "Any association of three or more people that is organised under a name, sign or symbol and which has an ideology, statements or physical actions that attack individuals based on characteristics, including race, religious affiliation, nationality, ethnicity, gender, sex, sexual orientation, serious disease or disability."[53]

ARTICLE 19 has the following concerns with this section:

- The definition of **'hate organisations'** is incredibly broad and could arguably include some political parties, though the threshold that would need to be reached for such organisations to be banned remains unclear.

- The term 'attack' remains unclear and there is no mention of intent to harm particular groups. Coupled with a broad ban on "content that praises any of the above organisations or individuals or any acts committed by the above organisations or individuals" and "co-ordination of support for any of the above organisations or individuals or any acts committed

[52] Facebook Community Standards, Dangerous Individuals and Organisations.
[53] *Ibid.*

by the above organisations or individuals," it seems inevitable that content considered legitimate under international law will get caught up within these definitions.

Exceptions

While the "Dangerous individuals and organisations" does not make any allowance for exceptions, the 'hate speech' section explains that sharing content for the purposes of raising awareness and educate others is allowed. However, users are expected to make clear their intent or the content might be removed. Humour and social commentary are also allowed. Whilst the exceptions are welcome, they are unduly limited. In particular, a requirement to state intent to educate others sets a high threshold for the exceptions to apply. It seems unlikely and unrealistic to expect users to state their intent explicitly whenever they comment or joke about an issue. Moreover, people may have a different sense of humour. However, there is no guidance (e.g in the form of specific examples) from Facebook on how it applies these standards in practice.[54]

Recommendations:

- Facebook's definition of 'hate speech' should be more closely aligned with international standards on freedom of expression, including by differentiating between different types of 'hate speech;'
- Facebook should provide case studies or more detailed examples of the way in which it applies its policies on 'hate speech;'
- The definition of 'hate organisations' and 'attack' should be narrowed, including by adding a requirement of intent to harm particular groups;
- Failure to express intent in relation to 'hate speech' (whether educational or otherwise) should not lead to automatic removal of that content in practice.

Extremism/Terrorism

Facebook primarily deals with 'terrorist' content under the 'dangerous organisations' section.

Accordingly, the Community Standards provide that Facebook does not allow on its platform any organisations or individuals that are engaged in "terrorist organisations, organised hate, mass or serial murders, human trafficking, organised violence or criminal activity."[55] The Community Standards further explain that Facebook removes content that "expresses support or praise for groups, leaders or individuals involved in these activities." The key terms are defined as follows:

- 'Terrorist organisations' are defined as "any non-governmental organisation that engages in premeditated acts of violence against persons or property to intimidate a civilian population, government or international organisation in order to achieve a political, religious or ideological aim;"

- 'Terrorist' is "a member of a terrorist organisation or any person who commits a terrorist act is considered a terrorist;"

- 'Terrorist act' is defined as "a premeditated act of violence against persons or property carried out by a non-government actor to intimidate a civilian population, government or international organisation in order to achieve a political, religious or ideological aim."

[54] More recently, Facebook published a blog post giving some examples in its Hard Questions Series, Why do you leave up some posts but take down others, 24 April 2018. However, the examples are very limited and do not form part of Facebook's community guidelines. As such, they are unlikely to be read by its users.
[55] *Op.cit.*

Facebook denies a free speech platform to individuals or organisations engaged in terrorist activity. While it is understandable and not necessarily an unreasonable restriction on freedom of expression in and of itself, a key difficulty is the lack of agreed definition of terrorism at international level.[56] Although Facebook's definition of terrorism contains a number of positive elements (such as an explicit reference to 'premeditated acts of violence,' the purpose of 'intimidating a civilian population, government, or international organisation' and narrowing the motivations of these actors to 'political, religious or ideological aims'), it could be more narrowly defined. 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." [57]

Moreover, Facebook's definition leaves several questions unanswered, for instance whether an organisation launching a cyber-attack on critical infrastructure would fall within its definition of terrorist activity. This is compounded by the lack of explicit examples being given of who or what falls in the definition of 'terrorist organisation.' For instance, Facebook recently banned from its platform the Arakan Rohingya Salvation Army (ARSA), leading to the deletion of posts, on the basis of their support for ARSA, by Rohingya civilians fleeing alleged ethnic cleansing operations by the military.[58] Facebook explained that this was an error. However, better guidance might have helped prevent this problem.

More generally, it is unclear how Facebook deals with individuals designated as 'terrorist' by certain governments but who may otherwise be regarded as freedom fighters. While Facebook has stated in conferences that it complies with the US State Department list of designated terrorist groups, this is not made explicit in the community standards. This can also 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 Facebook 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.

Moreover, Facebook's policy of banning content that "expresses support" or "praise" for those groups, leaders or individuals involved in these activities is overbroad and inconsistent with international standards on freedom of expression.[59] In particular, 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. For content to amount to incitement to terrorism, there should be an objective risk that the act incited will be committed and intent that the message at issue incites the commission of a terrorist act.[60]

Finally, it is worth noting that Facebook may ban terrorist content under separate rules, namely those on **'graphic violence.'**[61] The Community Standards provide that Facebook removes content that "glorifies violence or celebrates the suffering or humiliation of others because it may create an environment that discourages participation." At the same time, Facebook notes that "people value the ability to discuss important issues such as human rights abuse or acts of terrorism." As such, it allows graphic content "to help people raise awareness about issues" but adds a "warning label to especially graphic or violent content," among other things to prevent under-18s to have access to that content. This includes, among other things, imagery featuring mutilated people in a medical setting; videos of self-immolation when that action is a form of political speech or

[56] See e.g. UN Special Rapporteur on Counter-Terrorism, A/HRC/16/51, 2010.
[57] *Ibid.*
[58] See Dhaka Tribune, Facebook brands ARSA a dangerous organization, bans posts, 20 September 2017.
[59] See A/66/290, *op. cit.* See also, ARTICLE 19, The Johannesburg Principles on National Security, Freedom of Expression and Access to Information, 1996.
[60] See A/66/290, *op. cit.*
[61] Facebook Community Standards, Graphic violence.

newsworthy; photos of wounded or dead people; videos of child or animal abuse; and videos that show acts of torture committed against a person or people.

ARTICLE 19 notes that the Community Standards on this topic are relatively specific, though it is difficult to know how they are applied in practice in the absence of case studies. For instance, the Community Standards suggest that Facebook would allow beheading videos when shared for journalistic or educational purposes, despite the fact that they may have been disseminated by terrorist groups in the first place. This would be in keeping with international standards on freedom of expression, but there is no data available to confirm that this is indeed the case, for instance in Facebook's Transparency Report.

Recommendations:

- Facebook should align its definition of terrorism with that recommended by the UN Special Rapporteur on counter-terrorism. In particular, it should avoid the use of vague terms such as 'praise,' 'express support,' 'glorification' or 'promotion;'
- Facebook should give examples of organisations falling within the definition of 'terrorist organisation'. 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.
- Facebook should provide case studies explaining how it applies its standards in practice (e.g. on beheading videos).

Privacy and morality-based restrictions

Facebook restricts several types of content, which are relevant to the protection of privacy and public morals, including child nudity and sexual exploitation of children,[62] sexual exploitation of adults,[63] bullying,[64] harassment,[65] privacy breaches and 'image privacy rights'[66] under the 'safety' heading; cruel and insensitive content,[67] and adult nudity and sexual activity [68] under its 'objectionable content' heading; and "credible threats of violence" under its 'violence and criminal behaviour heading.'[69]

In this section, ARTICLE 19 focuses on (i) nudity, (ii) bullying, harassment, and cruel and insensitive content; (iii) credible threats of violence and (iv) privacy.[70]

Nudity

Facebook's approach to adult nudity has long been controversial. The Community Standards explain that Facebook restricts the display of nudity or sexual activity "because some people in our community may be sensitive to this type of content."[71] It goes on to state that it:

> "default[s] to removing sexual imagery to prevent the sharing of non-consensual or underage content [and that] restrictions on the display of sexual activity also apply to

[62] Facebook Community Standards, Child nudity and sexual exploitation of children.
[63] Facebook Community Standards, Sexual exploitation of adults.
[64] Facebook Community Standards, Bullying.
[65] Facebook Community Standards, Harassment.
[66] Facebook Community Standards, Privacy breaches and image privacy rights.
[67] Facebook Community Standards, Cruel and insensitive.
[68] Facebook Community Standards, Adult nudity and sexual activity.
[69] Facebook Community Standards, Credible violence.
[70] See, for instance, ARTICLE 19, Global Principles on the Protection of Freedom of Expression and Privacy, 2017. The principles seek to provide a progressive interpretation of international and comparative standards in this area.
[71] Adult nudity, *op. cit.*

digitally created content unless it is posted for educational, humorous or satirical purposes."

In allowing exceptions to the rule, Facebook explains that it understands that "nudity can be shared for a variety of reasons, including as a form of protest, to raise awareness about a cause or for educational or medical reasons." As such, it makes allowance for this type of content "where such intent is clear." As an example, it explains "while we restrict some images of female breasts that include the nipple, we allow other images, including those depicting acts of protest, women actively engaged in breastfeeding and photos of post-mastectomy scarring." The company also allows "photographs of paintings, sculptures and other art that depicts nude figures." A further section of the policy goes on to detail the company's definition of 'nudity,' 'sexual activity' and 'sexual intercourse.'[72]

ARTICLE 19 generally welcomes Facebook's more detailed policies on nudity, including its explanation about the rationale for them, for example, that it wishes to prevent the sharing of non-consensual images. Nonetheless, it appears that these restrictions remain chiefly motivated by a desire to 'protect' users generally from seeing certain forms of sexualised content. A key concern in this area is that there is still a lack of clarity as to how these 'morality-based' terms may be enforced discriminatorily against sexual expression by women and/or lesbian, gay, bisexual, and transgender (LGBT) persons. Decisions to remove such content often appear to be inconsistent with the treatment of analogous expression by cis-gendered men or heterosexual people.

Bullying, harassment, and cruel and insensitive content

Facebook also has dedicated policies to deal with bullying, harassment and cruel and insensitive content. It makes clear that neither bullying nor harassment is tolerated on the platform.

Bullying content is removed insofar as it "purposefully targets private individuals with the intention of degrading or shaming them."[73] Among other things, users are invited not to post the following types of content:

- Content about another private individual that reflects claims about sexual activity, degrading physical descriptions about or ranking individuals on physical appearance or personality, threats of sexual touching, sexualised text targeting another individual or physical bullying where the context further degrades the individual;

- Images that have been edited to target and demean an individual, including by highlighting specific physical characteristics or threatening violence in text or with imagery; and

- Content that specifies an individual as the target of statements of intent to commit violence, calls for action of violence, statements advocating violence, aspirational and conditional statements of violence, 'physical bullying.'

- In addition, Facebook indicates that it may remove Pages or Groups that are dedicated to attacking individuals, e.g. by cursing, negative character claims or negative ability claims. When minors are involved, this content is removed, alongside other types of content, such as attacks on minors by negative physical description.

ARTICLE 19 notes that 'bullying' is an intrinsically broad concept. For this reason, it is not usually defined in legislation.[74] We further note that Facebook's own definition may sometimes reflect existing offences, such as threats of violence, which are perfectly compatible with international

[72] *Ibid.*
[73] Bullying, *op.cit.*
[74] See for instance the Nova Scotia Cyber Safety Act 2013, which sought define bullying. The Act was struck down by the Supreme Court of Nova Scotia; see *Crouch v. Snell*, 2015 NSSC 340, p. 47 ff.

standards on freedom of expression. However, Facebook's understanding of 'bullying' goes much beyond that to include merely offensive and distasteful comments, such as degrading physical descriptions or ranking of individuals based on personality traits or physical attractiveness. It also covers "cursing, negative character claims or negative ability claims." It is further unclear whether any exceptions are applied, for instance when 'negative character claims' are made about public officials, politicians and the like. In other words, Facebook's definition of bullying is particularly broad and falls below international standards on freedom of expression.

By contrast, **harassment** is not expressly defined in Facebook's community standards, though the company provides several examples of prohibited content.[75] Contrary to its bullying policy, Facebook clearly states that context and intent matter, and that it allows people to share and re-share posts "if it is clear that something was shared in order to condemn or draw attention to harassment."[76] The main distinction between harassment and bullying seems to lie in removing content when it involves "repeatedly contacting a single person despite that person's clear desire and action to prevent that contact" or "repeatedly contacting large numbers of people with no prior solicitation." At the same time, 'harassment' remains broadly defined and encompasses messages in breach of its bullying policy. Other examples of 'harassment' include "claims that a victim of a violent tragedy is lying about being a victim, acting/pretending to be a victim of a verified event, or otherwise is paid or employed to mislead people about their role in the event when sent directly to a survivor and/or immediate family member of a survivor or victim," or "targeting anyone maliciously," including public figures, by "threatening any participant in public discourse with violence in an attempt to intimidate or silence them."[77]

ARTICLE 19 generally welcomes Facebook's detailed policy on harassment. However, we note that Facebook's understanding of harassment goes much beyond traditional legal definitions of that concept, particularly since it also includes breaches of its bullying policy. We further note that the policy does not explain how it deals with claims of 'harassment', which in reality merely amount to unwanted criticisms. This is particularly relevant in the case of politician's pages or groups. For instance, it is unclear whether politicians with a public profile should be permitted to block individuals when they receive negative feedback about their positions or character as opposed to credible threats of violence. Equally, it is unclear whether such content should be removed on the basis that it constitutes 'harassment.'

Finally, these policies are consolidated in a section on **'cruel and insensitive' content**, which states that the company has "higher expectations for content that we call cruel and insensitive, which we define as content that targets victims of serious physical or emotional harm." As such users should not post content that depicts real people and "mocks their implied or actual serious physical injuries, disease or disability, non-consensual sexual touching or premature death."

In ARTICLE 19's view, these restrictions on content are particularly broad. Though understandable, they fall below international standards on freedom of expression.

Credible threats of violence

Facebook may also remove credible threats of violence against individuals.[78] In particular, it explains that in doing so, it considers factors such as "language, context and details in order to distinguish casual statements from content that constitutes a credible threat to public or personal safety." It also considers factors such as a person's public visibility and vulnerability," and any

[75] Harassment, *op.cit.*
[76] However, Facebook states that its bullying policy does not apply to public figures in order to allow discourse, "which often includes critical discussion of people who are featured in the news or who have a large public audience."
[77] For more examples, see Facebook's community standards on harassment, *op. cit.*
[78] Credible violence, *op.cit.*

additional information such as threats mentioning a target and a bounty/demand for payment, a reference or image of a specific weapon, details about location, timing and method, etc.[79]

ARTICLE 19 notes that Facebook's policy on credible threats of violence is relatively detailed, it makes reference to 'statements of intent,' which is positive, and lists a number of factors it takes into account in assessing the credibility of threats. We note however, that its section on 'credible violence' overlaps with its policy on 'hate speech,' which could create some confusion, particularly in circumstances where Facebook does not explain the distinction between the two. We also note that 'outing' an individual is equated with 'violence'. In our view, this is overbroad and would be better placed under Facebook's 'privacy breaches' heading or as part of its harassment policy given the potential real-word impact that such outing might have.

Privacy

The section on privacy breaches in the Facebook's Community Standards is relatively wide-ranging:

- It bans the posting of several categories of private information that facilitate identity theft such as national or school IDs, social security information, bank account or credit card numbers, personal medical records, private phone numbers or addresses and email;

- It further enjoins users not to post "except in limited cases of newsworthiness, content claimed or confirmed to come from a hacked source, regardless of whether the affected person is a public figure or a private individual;"

- It also prohibits "Content that includes photographs that display the external view of private residences if the following conditions apply: (i) the residence is a single-family home, or the resident's unit number is identified in the image/caption; (ii) the town/city or area is identified; (iii) a resident is mentioned or depicted; and (iv) that same resident objects to the exposure of their private residence.

- Facebook also warns that certain photos or videos may be removed where the person depicted in the image is (i) a minor under thirteen years old, and the content was reported by the minor or a parent or legal guardian; (ii) a minor between thirteen and eighteen years old, and the content was reported by the minor; (iii) an adult, where the content was reported by the adult from outside the United States of America and applicable law may provide rights to removal; (iv) any person who is incapacitated and unable to report the content on their own.

In ARTICLE 19's view, these rules appear to be broadly in keeping with data protection and common-sense rules for the protection of the right to privacy. At the same time, Facebook's policy 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 make reference to the criteria developed, *inter alia,* in the Global Principles on the Protection of Freedom of Expression and Privacy.[80]

ARTICLE 19 concludes that although Facebook's Community Standards are generally drafted in relatively plain language, they tend to ban broader categories of content than those permitted under international standards on freedom of expression. Beyond 'hate speech,' this is especially true of content considered 'abusive' 'harmful' 'negative' or which constitutes 'bullying,' none of which would normally be prohibited by law.

At the same time, we recognise that companies might legitimately restrict access to some lawful content because of the type of service they want to provide, i.e. creating a safe environment

[79] *Ibid.*
[80] *Op. cit.*

suitable for a wide audience (see for instance, Facebook's position on nudity). The main problem is that community guidelines are drafted in broad terms giving companies flexibility to interpret them according to their own needs. This results in inconsistent and sometimes apparently biased outcomes. In the absence of more concrete examples being given of how the guidelines are applied, it is difficult to know what content actually gets removed from these platforms.

Recommendations:

- Facebook should strive to narrow its definitions of bullying and harassment in order to prevent legitimate content from being removed;
- Facebook should ensure that the definition of bullying and harassment remain distinct;
- Facebook should explain in more detail the relationship between 'threats,' 'harassment,' and 'online abuse'/'bullying,' and distinguish this from 'offensive content' (which should not be limited as such);
- Facebook 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;
- Facebook should state more clearly that offensive content will not be taken down as a matter of principle unless it violates other rules.
- Facebook 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 make reference to the criteria developed, *inter alia,* in the Global Principles on the Protection of Freedom of Expression and Privacy.

'Fake news'

Facebook does not allow the use of inaccurate or misleading information in order to collect likes, followers or shares as part of its anti-spam policy.[81] It also removes profiles that impersonate other people.[82]

In response to the challenge of 'fake news', Facebook started working with fact-checking organisations in 2016 in order to put in place a 'fake news' labelling system.[83] Under this new initiative, readers were able to alert Facebook that a story might be false. If enough people reported that story as fake, it was then sent to trusted third-party fact-checkers. If the story was deemed unreliable, it became publicly flagged as "disputed by third-party fact checkers" and a warning appears when users decide to share it.[84] Following criticism that this system was not effective,[85] Facebook decided to replace public flags with "related articles" to provide more context to stories reported as inaccurate.[86]

In 2018, Facebook announced that it would seek to tackle 'fake news' and 'clickbait' by no longer prioritising Pages or public content in its News Feed. Rather, it would prioritise content published by family and friends[87] or content rated as trustworthy by the Facebook community.[88] More emphasis would be placed on news that people find informative and news that is relevant to people's local community.[89] Meanwhile, Facebook has explained that it has been hunting down 'fake' accounts and worked with government and civil society partners to defend its platform from

[81] Facebook Community Standards, Spam.
[82] Facebook Community Standards, Misrepresentation. Facebook justifies these rules and its real-name policy by reference to the need for trust and accountability on its platform.
83 The Guardian, Facebook to begin flagging fake news in response to mounting criticism, 15 December 2016.
[84] *Ibid.*
85 The Guardian, Facebook promised to tackle fake news. But the evidence shows it's not working, 16 May 2017.
86 Facebook Newsroom, Replacing Disputed Flags With Related Articles, 20 December 2017.
[87] Facebook Newsroom, Bringing People Closer Together, 11 January 2018.
88 Facebook Newsroom, Helping Ensure News on Facebook Is From Trusted Sources, 19 January 2018.
[89] *Ibid.*

"malicious interference," particularly during elections.[90] Also, it is strengthening its enforcement of its ad policies[91] and continues to create new tools to help its users better understand the context of the articles in its News Feed.[92]

In the 2018 iteration of the Community Standards, Facebook states that "it doesn't remove false news," chiefly in order not to stifle public discourse, but that it "significantly reduces its distribution by showing it lower in News Feed."[93] It does so, among other things, "using various signals, including feedback from our community, to inform a machine learning model that predicts which stories may be false."[94]

ARTICLE 19 notes that the absence of an outright ban on misinformation or 'fake news' in Facebook's terms of service is to be welcomed. Equally, its voluntary initiatives aimed at identifying 'fake news' in cooperation with fact-checkers are a positive step, insofar as they do not generally involve the removal of information.

Nonetheless, voluntary initiatives aimed at identifying 'fake news' are work in progress. For instance, it remains unclear whether flagging or offering related content is effective[95] As such, these initiatives could be further improved. In particular, they should include a charter of ethics comparable to the highest professional standards of journalism, be as open and transparent as possible and involve a wide range of stakeholders in order to ensure that internet users receive a real diversity of opinions and ideas and are able to identify misinformation.[96]

By contrast, Facebook's decision to prioritise content posted by friends at the expense of public content seems to suggest that it would rather shirk responsibility for the quality of the information on its network than engage with content publishers. It is highly unclear whether this type of initiative would contribute to better quality information and is more likely to create difficulties when, as already noted by the UN Special Rapporteur on FoE, community-trusted news sources come into conflict with government-trusted ones.[97]

More concerning, however, is the lack of transparency surrounding the 'tweaking' of the algorithms of companies such as Facebook in order to produce 'reliable' results or 'good' content. In particular, there is a real risk that the content of small media companies might become less visible by pushing them further down the list of 'recommended' content.[98] This raises significant issues for media pluralism, diversity and competition. Moreover, the closing of 'fake' accounts on the basis of 'suspicious' activity raises questions about the more detailed criteria and systems used by Facebook in order to take such decisions. There is currently little to no information available to ensure that Facebook does not close accounts by mistake and what redress is available when errors are made.

Recommendations:

- Facebook should improve its voluntary initiatives aimed at tackling 'fake news' by including a charter of ethics comparable to the highest professional standards of journalism and involving a wide range of stakeholders;

90 Facebook Newsroom, Update on German Elections, 27 September 2017.
91 Facebook Newsroom, Improving Enforcement and Transparency of Ads on Facebook, 2 October 2017.
92 Facebook Newsroom, New Test to Provide Context About Article, 5 October 2017/
93 Facebook Community Standards, False News.
94 *Ibid.*
95 See e.g. Poynter, It's been a year since Facebook partnered with fact-checkers. How's it going?, 15 December 2017.
96 See ARTICLE 19, Social media and fake news from a free speech perspective, 25 November 2016. See also ARTICLE 19's policy on 'fake news,' forthcoming.
97 The New York Times, Facebook to Let Users Rank Credibility of News, 19 January 2018.
98 See e.g. the impact of Facebook's changes to its news feed in some test countries, including Cambodia, The Forbes, Facebook's Explore Feed Experiment Is Already Crushing Cambodia's Businesses, 2 November 2017.

- Facebook should explain in more detail how its algorithms detect 'fake' accounts or produce more 'reliable' results, including by listing the criteria on the basis of which these algorithms operate.

Content removal processes

Facebook provides various reporting mechanisms on its platform, from reporting particular accounts, Pages, posts and so on,[99] to communication tools that allow users to request that other users take content down ('social reporting').[100] In addition, Facebook relies on algorithms to filter out certain types of content and uses a trusted flagger system to fast-track reports of violations of its community standards. Facebook also places significant emphasis on other tools to address abuse such as hiding news feeds or blocking or unfriending individuals.[101]

At the same time, Facebook does not appear to give any reasons to users for the restrictions it applies to their accounts in response to a report.[102] Nor does it seem to provide for any clear appeals or review mechanism of its decisions.[103] ARTICLE 19 finds that while social reporting and Facebook's emphasis on other tools to prevent exposure to undesirable content are welcome, the lack of clear appeals mechanism in relation to wrongful content removals or other sanctions is a fundamental flaw in its internal system. The absence of reasons for content restrictions is also inconsistent with due process safeguards.

More generally, ARTICLE 19 notes that Facebook, like other similar companies, increasingly relies on algorithms and a 'trusted flagger' system to speed up its content removal process, sometimes preventing content from being published altogether in the first place. We believe that it is imperative for Facebook to show leadership in this area and at least be more transparent in relation to its use of these tools.

- The lack of transparency in relation to the use of algorithms in order to detect particular types of content, such as 'extremism' or 'hate speech,' means that companies such as Facebook are more likely to be prone to bias. It is also unclear how algorithms can be trained to take into account free speech concerns (i.e. context), if at all;

- Equally, Facebook does not currently provide any meaningful information about the 'trusted flagger' system and the extent to which content flagged by such flaggers are subject to adequate review or are automatically removed. Although the trusted flagger system may contribute to better quality notices, it should in no way be taken as equivalent to an impartial or independent assessment of the content at issue. Trusted flaggers are often identified due to their expertise on the impact of certain types of content, whether copyright, terrorism-related content, or 'hate speech,' and their proximity to victims of such speech but not on the basis of having freedom of expression expertise. They are therefore not necessarily well-placed to make impartial assessments of whether restricting the content at issue is consistent with international human rights law.

Finally, ARTICLE 19 notes that Facebook may sometimes be required to put in place national contact points as a matter of law.[104] Equally, it may also do so voluntarily.[105] This is a matter of

[99] Facebook Help Centre, How to Report Things
[100] Facebook Help Centre, What is Social reporting.
[101] Facebook Help Centre, How do I report inappropriate or abusive things on Facebook (example: nudity, hate speech, threats)?
[102] Facebook Help Centre, Report Something.
103 See, however, Facebook's April 2018 announcement that it would be strengthening its appeals process, see Facebook Newsroom, Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process, 24April 2018.
[104] See e.g. ARTICLE 19, Germany: The Act to Improve Enforcement of the Law in Social Networks, August 20187.

concern, particularly in countries with governments with a poor record on the protection of freedom of expression. In particular, national points of contact may facilitate the removal of legitimate content under international human rights law. As such, we believe that Facebook should refrain from voluntarily putting in place national contact points, particularly in countries with governments with a poor record on the protection of freedom of expression.

Recommendations:

- Facebook 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;
- Facebook should be more transparent about its use of algorithms to detect various types of content, such as 'terrorist' videos, 'fake' accounts or 'hate speech;'
- Facebook should provide information about its trusted flagger system, including by identifying members of the scheme and the criteria being applied to join it.
- Facebook should refrain from putting in place contact points in countries with a poor record on the protection of freedom of expression.

Sanctions

Facebook provides that "the consequences of breaching our Community Standards vary depending on the severity of the breach and a person's history on Facebook. For instance, we may warn someone for a first breach, but if they continue to breach our policies, we may restrict their ability to post on Facebook or disable their profile."[106]

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 Facebook, like many other similar companies, appears to be increasingly applying country filters so that the promise of free expression 'beyond borders' is rapidly evaporating.[107]

Recommendations:

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

Real-name and identification policies

Facebook requires its users to use their real name when posting content on its platform.

In our view, this is inconsistent with international standards on free expression and privacy:

- The use of real-name registration by social media platforms and news sites as a prerequisite to using their services can have a negative impact on the rights to privacy and freedom of expression, particularly for minority or vulnerable groups, who might be prevented from asserting their sense of identity;

- Whilst real-name policies are usually presented as an effective tool against internet trolling, fostering a culture of mutual respect between internet users, the disadvantages of real-name

[105] The New York Times, Facebook Faces a New World as Officials Rein in a Wild Web, 17 September 2017.
[106] The Community Standards, *op.cit.*
[107] See e.g. EFF's report, Who has your back? Censorship Edition, 2018.

policies outweigh their benefits. In particular, anonymity is vital to protect children, victims of crime, individuals from minority groups and other vulnerable groups from being targeted by criminals or other malevolent third parties who may abuse real-name policies. In this sense, anonymity is as much about online safety as self-expression.

We are further concerned that real-name policies, such as Facebook's, are often accompanied by a requirement to provide identification. For instance, Facebook lists the different types of IDs it accepts in order to confirm its users' identity.[108] This, in our view, raises serious concerns over data protection, given that many such demands require users to provide a considerable amount of sensitive personal data as a means to verify their identity. Even if a company were to delete such data immediately, the very existence of the company's policy could put users at risk in certain countries. In particular, governments could more easily track down dissidents since they would already be identified by their Facebook account.

Recommendations:

- Facebook should stop requiring its users to use their real-name;
- Facebook should not require users to prove their identity.

[108] Facebook The Help Centre, What types of ID does Facebook accept?

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 http://www.article19.org/resources.php/legal.

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.