ARTICLE 19

YouTube Community Guidelines

September 2018

Legal analysis

Executive summary

In this analysis, ARTICLE 19 reviews the compatibility of YouTube's Community Guidelines with international standards on freedom of expression. Our analysis is based on the YouTube Community Guidelines as accessed in August 2018.

YouTube's Community Guidelines are divided into several sections, including nudity or sexual content; hateful content; harassment and cyberbullying; threats; privacy; child safety; harmful and dangerous content; violent or graphic content; spam, misleading metadata and scams; copyright; impersonation; and additional policies. The latest version of the Community Guidelines is generally easy to navigate and appears to contain more detailed information than in the past, such as a new Harassment and Cyberbullying policy, which is a welcome development. However, our analysis shows that YouTube's Community Guidelines fall below international standards on freedom of expression in a number of areas. This includes YouTube's content policies on 'terrorism' and 'harassment and cyberbullying,' and its complaint mechanisms.

ARTICLE 19 encourages YouTube to bring its Community 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. YouTube should set out in more detail the factors it relies on in assessing 'hate speech.' In addition, it should provide case studies or more detailed examples of the way in which it applies its policies on 'hate speech';

2. YouTube's 'hate speech' policy could be further developed so that it would differentiate between different types of 'hate speech';

3. YouTube should align its definition of terrorism and incitement to terrorism with that recommended by the UN Special Rapporteur on counter-terrorism and human rights. In particular, it should avoid the use of vague terms such as 'celebrate' or 'promotion' of terrorism;

4. YouTube should give examples of organisations falling within the definition of 'terrorist' organisations. 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;

5. YouTube should provide case studies explaining how it applies its 'terrorism' standards in practice;

6. YouTube should elaborate its policy on nudity and sexual content, including giving clearer examples of the types of content that are likely to be removed under the policy;

7. YouTube should explain what constitutes sufficient information for the purposes of providing context under its Nudity and Graphic Content policies. In practice, YouTube should not place too high a burden on users to provide contextual information. In particular, the absence of contextual information should not lead to automatic removal of content that may otherwise be legitimate under international standards on freedom of expression;

8. YouTube should define what constitutes a "malicious" attack and explain what factors are taken into account to distinguish "offensive" from "abusive" content. It should also consider adding

a reference to causing "alarm or distress" in its definition of harassment. Harassment should be more clearly distinguished from bullying;

9. YouTube should provide exceptions to its Harassment and Cyberbullying policies so as to protect freedom of expression, in particular legitimate criticisms that may be deemed offensive by the individuals concerned;

10. YouTube should provide examples or case studies of how its Harassment and Cyberbullying policy is applied in practice;

11. YouTube to explain the relationship between its Harassment and Cyberbullying policy and its Hate Speech policy where appropriate;

12. YouTube's policy on Threats should make clear that threats of violence must at least be credible;

13. YouTube should clarify what falls within "encouragement" of "dangerous" or "illegal activities" in its Harmful and Dangerous Content policies;

14. YouTube should provide more examples of the way in which it applies its policies on Threats and Harmful and Dangerous Content;

15. YouTube should make reference to the more detailed criteria developed, *inter alia,* in Principle 12 of the Global Principles on the Protection of Freedom of Expression and Privacy as part of its assessment of privacy complaints. It should also provide examples or case studies of the way in which it applies those standards in practice;

16. YouTube should explain more clearly how its policies on spam and "deceptive practices" are related to the broader policy debates on 'fake news' or the dissemination of false information;

17. YouTube should be more transparent about the extent to which it might remove "false information" or "fake accounts" in practice;

18. YouTube should explain what it considers to be an "authoritative" source of news and how its algorithm promotes such sources;

19. YouTube should ensure that its appeals process complies with the Manila Principles on Intermediary Liability, particularly as regards counter-notices and the giving of reasons for actions taken;

20. YouTube should provide disaggregated data on the number of appeals filed and their outcome in its Transparency Report;

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

22. YouTube should provide more details about the members of its Trusted Flagger Program.

Table of contents

Introduction

In this analysis, ARTICLE 19 reviews the YouTube Community Guidelines on content (the August 2018 version).

Since its inception in 2005, YouTube has grown into a multi-billion-dollar company. YouTube has over one billion users who watch a billion hours of video each day.[1] It is 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 central player in the regulation and moderation of online content.

Over the years, YouTube has amended its Community Guidelines to make them more accessible to its users. YouTube's Community Guidelines are now divided into the following sections:

- Nudity or sexual content
- Hateful content
- Harassment and cyberbullying
- Threats
- Privacy
- Child safety
- Harmful and dangerous content
- Violent or graphic content
- Spam, misleading metadata and scams
- Copyright
- Impersonation
- Additional policies

ARTICLE 19 welcomes YouTube's ongoing efforts to clarify its rules on content moderation. The latest YouTube Community Guidelines Enforcement Transparency Report,[2] which among other things features content removals on the basis of automated detection or flagging system is another welcome development.

However, we find that the YouTube Community Guidelines fall short of international standards on freedom of expression in a number of areas. In particular, YouTube imposes restrictions on 'violent extremism' or 'terrorist' content that are inconsistent with applicable international standards. Several rules remain very broad in scope, leaving significant discretion to YouTube in their implementation. As such, they are highly likely to lead to inconsistent application, particularly in relation to 'Harassment and Cyberbullying.' Like other social media companies, YouTube should provide more case studies or detailed examples of its internal 'case-law'. It should also strive to bring its appeals processes in line with international standards on freedom of expression and due process.

ARTICLE 19 believes that social media companies, including YouTube, 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 YouTube, 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

[1] See YouTube For Press.
[2] YouTube Community Guidelines Enforcement Report; the report is issued quarterly, the first report was issued in April 2018.

analyse the YouTube Community Guidelines in some key areas, focusing on 'hate speech,' 'terrorist' content, privacy and morality-based restrictions on content, and 'fake news.' We also examine YouTube's content removal processes and sanctions. Each section contains recommendations on how to bring the YouTube Community Guidelines in line with international standards on freedom of expression.

International human rights standards

ARTICLE 19's comments on YouTube Community Guidelines 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),[3] 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.[4]

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.[5]

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.[6]

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.

[3] 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).
[4] See HR Committee, General Comment No. 34 on Article 19: Freedoms of opinion and expression, CCPR/C/GC/34, 12 September 2011, para 11.
[5] HR Committee, *Belichkin v. Belarus*, Comm. No. 1022/2001, U.N. Doc. CCPR/C/85/D/1022/2001 (2005).
[6] 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.[7]

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.[8] 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.[9]

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.[10] 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.[11]

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.[12] 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:[13]
  - 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

[7] 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.
[8] The Report of the Special Rapporteur on FOE, 16 May 2011, A/HRC/17/27, para 43.
[9] *Ibid.,* para 42.
[10] Report of the Special Rapporteur on FOE, 6 April 2018, A/HRC/38/35, para. 66.
[11] *Ibid.*, para. 69.
[12] 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.
[13] *Ibid.*, Principle 15.

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.

- 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.[14] 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.[15] 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.[16] 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.[17]

- In his **June 2016 Report to the Human Rights Council**,[18] 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"[19] 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."[20]

- In his **April 2018 Report to the Human Rights Council**,[21] 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.[22] 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.[23] Furthermore, the Special Rapporteur stressed the need that social media companies engage

[14] The May 2011 Report of the Special Rapporteur on FOE, *op.cit.*, para 45.
[15] *Ibid.* para 46.
[16] *Ibid.*, paras 47 and 76.
[17] *Ibid.*, paras 48 and 77.
[18] Report of the Special Rapporteur on FOE, 11 May 2016, A/HRC/32/38; para 40-44.
[19] Ibid.
[20] *Ibid.*, para 43.
[21] Report of the Special Rapporteur on FOE, 6 April 2018, A/HRC/38/35.
[22] *Ibid.*, paras. 41-43.
[23] *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."

more actively with civil society organisations[24] and open themselves up to public accountability mechanisms (such as a social media council)[25], 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[26] 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.[27] He also encouraged companies to publish transparency reports about government requests for user data or content removal;[28] challenge requests for content removal or requests for user data that may violate the law or internationally recognised human rights;[29] notify individuals affected by any measure restricting their freedom of expression and provide them with non-judicial remedies;[30] and take proactive protective measures to develop good business practices consistent with respect for human rights.[31]

- In the **2016 report on Standards for a Free, Open and Inclusive Internet**,[32] 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.[33] He also highlighted the need for companies' policies, operating procedures and practices to be transparent.[34]

- 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 to impose restrictions that violate States' human rights obligations.[35] 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.[36]

- 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.[37] Moreover, in its March 2018 **Recommendation on the roles and**

[24] *Ibid.*, para. 54.
[25] *Ibid.*, para. 58.
[26] 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."
[27] *Ibid.*, paras 111-112.
[28] *Ibid.*, para 113.
[29] *Ibid.*, para 114.
[30] *Ibid.*, para 115.
[31] *Ibid.*, para 116.
[32] OAS Special Rapporteur on FOE, Standards for a Free, Open and Inclusive Internet, 2016, paras 95-101.
[33] *Ibid.*, para 98.
[34] *Ibid.*, para 99.
[35] 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.
[36] *Ibid.*, p. 24.
[37] Committee of Ministers of Council of Europe, Recommendation CM/Rec (2012)4 of the Committee of Ministers

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.[38] 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.[39] In particular, they make clear that companies' content restriction practices must comply with the tests of necessity and proportionality under human rights law,[40] and that intermediaries should provide users with complaints mechanisms to review decisions to restrict content made on the basis of their content restriction policies.[41]

- 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).[42]

- Finally, the **Dynamic Coalition on Platform Responsibility** is currently seeking to develop standard Terms and Conditions in line with international human rights standards.[43]

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

---

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, 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).

[38] 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.

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

[40] *Ibid.,* Principle IV.

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

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

[43] 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.

with international standards.[44] Moreover, it is 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.[45]

- 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.[46] 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.[47] 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.[48] 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.[49]

- 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.[50] 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.[51] 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.[52]

---

[44] 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).
[45] *Ibid.*, para 2 i).
[46] 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).
[47] Ibid.
[48] *Ibid.*, para 4 b).
[49] *Ibid.*, para 4 c).
[50] 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.
[51] Ibid.
[52] 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 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, 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.[53]

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.[54]

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.[55] He also recommended that corporate actors reconsider their own policies that restrict encryption and anonymity (including through the use of pseudonyms).[56]

the promotion and protection of human rights and fundamental freedoms while countering terrorism Martin Scheinin, A/HRC/13/37, 28 December 2009.

[53] The May 2011 Report of the Special Rapporteur on FOE, *op.cit.*, para 53.

[54] *Ibid.*, para 84.

[55] Report of the Special Rapporteur FOE, A/HRC/29/32, 22 May 2015, para 60.

[56] Ibid.

Analysis of the YouTube Community Guidelines

'Hate speech'

The YouTube Community Guidelines contain a 'hate speech policy' section. It emphasises that YouTube encourages free speech and tries to defend users' right to express unpopular points of view, but does not "permit hate speech".[57]

'Hate speech' is defined as "content that promotes violence against or has the primary purpose of inciting hatred against individuals or groups based on certain attributes, such as race or ethnic origin, religion, disability, gender, age, veteran status, sexual orientation/gender identity".[58] YouTube also specifies that there is "a fine line between what is and what is not considered to be hate speech. For instance, it is generally okay to criticize a nation-state, but not okay to post malicious hateful comments about a group of people solely based on their race".[59] It further provides that users should "keep in mind that not everything that's mean or insulting is hate speech".[60]

ARTICLE 19 notes that the definition of 'hate speech' under YouTube's Hate Speech policy is close to – though wider than - the wording of Article 20(2) of the ICCPR, which requires States to prohibit the advocacy of discriminatory hatred that "constitutes *incitement* to discrimination, hostility or violence" (our emphasis). In particular, terms like "promotion" of violence are broader than "incitement". Nonetheless, YouTube should be commended for having a hate speech policy statement that is more closely aligned with international standards on freedom of expression than other social media companies such as Facebook or Twitter.

We also commend YouTube for the way in which it distinguishes the different types of content it moderates. For instance, identifying harassment and cyberbullying as a separate category from 'hate speech' or threats is helpful and conceptually clearer than content moderation categories used in other social media companies such as Twitter. Equally, YouTube's explanation that content should not be limited solely on the basis that it is offensive is to be welcomed.

At the same time, we believe that YouTube's 'hate speech' policy could be further improved. For instance, YouTube should clarify the meaning of terms such as "hateful" as well as the circumstances in which, for instance, an insult may amount to incitement to violence, hostility or discrimination, or amount to discriminatory threats or harassment. In particular, YouTube should provide more details about the factors it takes into account when determining whether a particular video or conduct amounts to "hate speech." In addition, YouTube should provide case studies or examples of how it applies those standards in practice.

Finally, we note that, 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:[61]

- 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;

[57] See YouTube, Hate Speech Policy.
[58] Ibid.
[59] Ibid.
[60] Ibid.
[61] For a full explanation of ARTICLE 19's policy on 'hate speech,' see Hate Speech Explained: A Toolkit, 2015.

- 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 YouTube, as a company, would not be expected to adopt the same types of measures as States, we believe that the above categories could further contribute to the elaboration of YouTube's response to 'hate speech.'

Recommendations:

- YouTube should set out in more detail the factors it relies on in assessing 'hate speech.' In addition, it should provide case studies or more detailed examples of the way in which it applies its policies on 'hate speech';
- YouTube's 'hate speech' policy could be further developed so that it would differentiate between different types of 'hate speech.'

Extremism/Terrorism

YouTube deals with terrorist content under the heading of "violent or graphic content".[62] YouTube makes clear that "[it] do[es] not permit terrorist organisations to use YouTube for any purpose, including recruitment". It goes on to explain that it "strictly prohibits content related to terrorism, such as content that promotes terrorist acts, incites violence or celebrates terrorist attacks". It further enjoins its users to provide enough information when posting content related to terrorism for an educational, documentary, scientific, or artistic purpose so that viewers understand the context. YouTube also warns that graphic or controversial footage may be subject to age-restrictions[63] or a warning screen.[64]

In ARTICLE 19's view, YouTube's ban on terrorist content is overly broad and inconsistent with international standards on freedom of expression:

- To begin with, we note that the YouTube ban on terrorist content goes far beyond "incitement to violence" or the "promotion" or "celebration" of terrorist acts or attacks since it is in principle applicable to all content simply "related" to terrorism.

- Secondly, rather than relying on the international definition of incitement to terrorism, YouTube uses vague and overbroad language such as the "promotion" of terrorist acts or the "celebration" of terrorist attacks, i.e. terms that 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.[65]

- Thirdly, YouTube does not define key terms such as "terrorist acts" or "violence." Furthermore, no reference is made to the question whether there is an objective risk that the act incited will be committed or the intent of the speaker that the message at issue will incite the commission of a terrorist act.[66]

---

[62] YouTube, Violent or Graphic Content Policies. We note that the wording August 2018 version is different from an earlier language; *c.f.,* ARTICLE 19, Sidestepping Rights: Regulating Speech by Contract, June 2018.
[63] YouTube Help, Age-restricted content.
[64] YouTube Help, Limited features for certain videos.
[65] See e.g. UN Special Rapporteur on freedom of expression, A/66/290, para. 34.
[66] See e.g. the model offence of incitement to terrorism in the UN Special Rapporteur on Counter-Terrorism's report, A/HRC/16/51, paras 29-32

We further note that, like other social media companies such as Facebook or Twitter, YouTube 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. This is compounded by the absence of a definition of 'terrorist activity' in YouTube's community standards. For instance, the UN Special Rapporteur on counter-terrorism and human rights 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". [67]

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

ARTICLE 19 welcomes the exceptions that YouTube seems to make in relation to 'terrorist' material posted for educational, documentary, scientific, or artistic purpose. However, YouTube does not explain what factors it takes into account in deciding whether a particular piece of 'terrorist' content falls within one of the permitted exceptions. In particular, YouTube does not explain what information users should provide to enable viewers to understand context. It is also unclear what constitutes "enough information" for that purpose.

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

Recommendations:

- YouTube should align its definition of terrorism and incitement to terrorism with that recommended by the UN Special Rapporteur on counter-terrorism and human rights. In particular, it should avoid the use of vague terms such as 'celebrate' or 'promotion' of terrorism;
- YouTube should give examples of organisations falling within the definition of terrorist organisations. 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;
- YouTube should provide case studies explaining how it applies its standards in practice.

Privacy and morality-based restrictions

YouTube, 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.
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.

[67] See A/HRC/16/51, *op. cit.*

Nudity and sexual content

YouTube's nudity and sexual content policies make clear that "sexually explicit content like pornography is not allowed".[68] YouTube goes on to explain that "videos containing fetish content will be removed or age-restricted depending on the severity of the act in question" and that "in most cases, violent, graphic or humiliating fetishes are not allowed to be shown on YouTube".[69] Equally, "if a video is intended to be sexually provocative, it is less likely to be acceptable for YouTube".[70]

By contrast, YouTube's nudity and sexual content policies explain that "a video that contains nudity or other sexual content may be allowed if the primary purpose is educational, documentary, scientific or artistic, and it isn't gratuitously graphic. For example, a documentary on breast cancer would be appropriate, but posting clips out of context from the same documentary might not be".[71] YouTube further encourages its users to provide context in the title and description of its users in order to help it and viewers to determine the primary purpose of the video.

Although separate from its nudity and sexual content policies, YouTube's policies on violent or graphic content also deserve to be mentioned here. In particular, YouTube bans violent or gory content that is primarily intended to be shocking, sensational or gratuitous.[72] YouTube explains, "if a video is particularly graphic or disturbing, it should be balanced with additional context and information".[73] As an example, YouTube explains that "a video by a citizen journalist which captures footage of protesters being beaten, uploaded with relevant information (date, location, context, etc.) would probably be allowed. However, posting the same footage without contextual or educational information may be considered gratuitous and may be removed from the site".[74] At the same time, YouTube warns that "in some cases, content may be so violent or shocking that no amount of context will allow that content to remain on our platforms."[75] As a halfway house, some graphic or violent content may be age-restricted.[76]

ARTICLE 19 notes that YouTube's policy on nudity and sexual content goes beyond international standards on freedom of expression that do not prohibit pornography as such but allow necessary and proportionate restrictions on access to that material. At the same time, we note that YouTube remains free to decide the type of service it wants to provide. Moreover, YouTube allows a number of exceptions to its nudity and sexual content policies, which is to be welcomed. Nonetheless, we note that YouTube's policies in this area are not particularly detailed so that users may not have a full understanding of what is allowed or not allowed on the platform. Moreover, YouTube appears to place too high a burden on users to provide sufficient contextual information, without explaining what sort of information might be relevant, for videos containing nudity or sexual content to be allowed. In practice, this means that YouTube has very wide discretion to remove legitimate content if users fail to provide enough context. The burden on users to provide sufficient information appears to be particularly high in relation to violent or graphic content, since YouTube acknowledges that videos of violence at a protest may be removed if insufficient information has been provided.

---

[68] See YouTube, Nudity and sexual content policies.
[69] *Ibid.*
[70] *Ibid.*
[71] *Ibid.*
[72] See YouTube, Violent or graphic content policies.
[73] *Ibid.*
[74] *Ibid.*
[75] *Ibid.*
[76] *Ibid.* More details are available on YouTube's age-restricted content policy page.

Harassment and cyberbullying

YouTube's policy on Harassment and Cyberbullying states YouTube's intent to protect its users from "malicious" harassment.[77] The policy explains that "in cases where harassment crosses the line into a malicious attack, it can be reported and the content will be removed. In other cases, users may be mildly annoying or petty and should simply be ignored".[78] YouTube goes on to list behaviour that falls within its definition of harassment, including:[79]

- Abusive videos, comments and messages
- Revealing someone's personal information, including sensitive personally identifiable information such as social security numbers, passport numbers or bank account numbers
- Maliciously recording someone without their consent
- Deliberately posting content in order to humiliate someone
- Making hurtful and negative comments/videos about another person
- Unwanted sexualisation, which encompasses sexual harassment or sexual bullying in any form
- Incitement to harass other users or creators

YouTube's Harassment and Cyberbullying policy also contains a number of tips to users on how to deal with unwanted behaviour. It advises users that "criticism and insults can escalate into more serious forms of harassment and cyberbullying. If specific threats are made against you and you feel unsafe, tell a trusted adult and report it to your local law enforcement agency".[80]

In addition, YouTube's Impersonation Policy states that "activities such as copying a user's channel layout, using a similar username or posing as another person in comments, emails or videos may be considered harassment" and lead to removal.[81]

ARTICLE 19 had previously noted that YouTube did not appear to have a standalone harassment policy.[82] As such, YouTube's Harassment and Cyberbullying policy is a welcome development. We note however that YouTube's definition of harassment is very broadly drafted and fails to define key terms. For example, YouTube does not explain what constitutes a "malicious" attack or what factors are taken into account to distinguish "offensive" from "abusive" content. We further note that harassment generally entails conduct causing "alarm or distress". However, these elements are not mentioned in the definition. Indeed, YouTube's definition of harassment sets a much lower threshold. In particular, YouTube's definition of harassment includes merely "hurtful" or "negative" comments. Legitimate criticisms may therefore be removed simply because of the hurt feelings of the individual who is the subject of those criticisms. Furthermore, "bullying" is undefined and it is unclear how it should be distinguished from harassment.

In light of the above, ARTICLE 19 believes that YouTube's Harassment and Cyberbullying policy is inconsistent with international standards on freedom of expression. In our view, YouTube should further define the terms outlined above and ensure that exceptions to the policy are clearly spelled out so as to protect freedom of expression. In addition, YouTube should provide examples or case studies of how its policy is applied in practice. Finally, we note that it might be helpful for YouTube to explain the relationship between its Harassment and Cyberbullying policy and its 'hate speech' policy.

---

[77] YouTube, Harassment and Cyberbullying policy.
[78] *Ibid.*
[79] *Ibid.*
[80] *Ibid.*
[81] YouTube, Impersonation policy.
[82] See ARTICLE 19, Sidestepping Rights: Regulating Speech by Contract, *op. cit.*

Threats

YouTube's policy on threats provides that "content that makes threats of serious physical harm against a specific individual or defined group of individuals will be removed".[83] It further explains that "people who threaten others may receive a strike on their account and their account may be terminated".

In addition, it is worth noting that YouTube's policy on harmful and dangerous content makes clear that it will remove content that intends to incite violence or encourage dangerous or illegal activities that have an inherent risk of serious physical harm or death.[84] YouTube goes on to explain that content that encourages dangerous or illegal activities include "instructional bomb making, choking games, hard drug use or other acts where serious injury may result".[85] As an exception, YouTube notes that "a video that depicts dangerous acts may be allowed if the primary purpose is educational, documentary, scientific or artistic (EDSA), and it isn't gratuitously graphic. For example, a news piece on the dangers of choking games would be appropriate, but posting clips out of context from the same documentary might not be".[86]

In ARTICLE 19's view, YouTube's policies on Threats and Harmful and Dangerous Content should be more narrowly drafted in order to be fully in line with international standards on freedom of expression. For instance, the policy on Threats does not specify that threats must be credible before YouTube will take enforcement action. In particular, the policy does not make a reference to a requirement of intent that the person to whom the threat has been issued would fear it would be carried out. In the absence of any examples, it is also unclear how YouTube interprets the policy. We note, for instance, that social media companies such as Twitter interpret threats as mere "wishes" for the serious physical harm, death, or disease of an individual, even though those may not be meant seriously. Equally, we believe that YouTube could further clarify what falls within "encouragement" of "dangerous" or "illegal activities". It is not entirely clear for instance whether this encompasses fictional depictions of such activities.

Privacy

YouTube's privacy policy can be found on its "Protecting your privacy" page[87] and in the YouTube Privacy Guidelines,[88] which provide a more detailed explanation of YouTube's privacy complaint process and the factors that it takes into account when evaluating privacy claims.

On its "Protecting your privacy" page, YouTube explains that "for content to be considered for removal, an individual must be uniquely identifiable",[89] In assessing if an individual is uniquely identifiable, YouTube considers the following factors:[90]

- Image or voice
- Full name
- Financial information
- Contact information
- Other personally identifiable information

YouTube goes on to explain that in assessing privacy complaints, it considers the public interest, newsworthiness and consent as factors in its final decision. The "Protecting your privacy" page further contains a number of tips on how users can protect their privacy on YouTube.

---

[83] See YouTube, Policy on Threats.
[84] See YouTube, Policies on Harmful or Dangerous Content.
[85] *Ibid.*
[86] *Ibid.*
[87] YouTube, Protecting your privacy .
[88] YouTube Privacy Guidelines.
[89] *Op. cit.*
[90] *Ibid.*

YouTube's Privacy Guidelines expand on the above by explaining that "for content to be considered for removal, an individual must be uniquely identifiable by image, voice, full name, government identification number, bank account number, contact information (e.g. home address, email address) or other uniquely identifiable information". YouTube goes on to explain that "to be considered uniquely identifiable, there must be enough information in the video that allows others to recognise you. Please note that just because you can identify yourself within the video, it does not mean you are uniquely identifiable to others. A first name without additional context or a fleeting image, for example, would not likely qualify as uniquely identifiable". In addition, YouTube reiterates that it takes "public interest, newsworthiness and consent into account when determining if content should be removed for a privacy violation". YouTube further reserves "the right to make the final determination of whether a violation of its privacy guidelines has occurred". In doing so, it makes clear that, although a video may not violate an individual's country's privacy laws, it may still violate YouTube's privacy guidelines. Finally, the YouTube Privacy Guidelines explain the Privacy Complaint Process[91] and how to report a privacy violation.

ARTICLE 19 notes that the YouTube Community Guidelines on privacy are broadly consistent with international standards on freedom of expression. However, the Privacy Guidelines do not provide detailed guidance on the types of factors that should be taken into account in balancing the rights to privacy and freedom of expression. In our view, YouTube should follow more closely the Global Principles on Protecting Freedom of Expression and Privacy in this respect.[92]

Recommendations:

- YouTube should elaborate its policy on nudity and sexual content, including giving clearer examples of the types of content that are likely to be removed under the policy;
- YouTube should explain what constitutes sufficient information for the purposes of providing context under its Nudity and Graphic content policies. In practice, YouTube should not place too high a burden on users to provide contextual information. In particular, the absence of contextual information should not lead to automatic removal of content that may otherwise be legitimate under international standards on freedom of expression;
- YouTube should define what constitutes a "malicious" attack and explain what factors are taken into account to distinguish "offensive" from "abusive" content. It should also consider adding a reference to causing "alarm or distress" in its definition of harassment. Harassment should be more clearly distinguished from bullying;
- YouTube should provide exceptions to its Harassment and Cyberbullying policies so as to protect freedom of expression, in particular legitimate criticisms that may be deemed offensive by the individuals concerned;
- YouTube should provide examples or case studies of how its Harassment and Cyberbullying policy is applied in practice;
- YouTube should explain the relationship between its Harassment and Cyberbullying policy and its Hate Speech policy where appropriate;
- YouTube's policy on Threats should make clear that threats of violence must at least be credible;
- YouTube should clarify what falls within "encouragement" of "dangerous" or "illegal activities" in its Harmful and Dangerous Content policies;
- YouTube should provide more examples of the way in which it applies its policies on Threats and Harmful and Dangerous Content;
- YouTube should make reference to the more detailed criteria developed, *inter alia,* in Principle 12 of the Global Principles on the Protection of Freedom of Expression and Privacy as part of its assessment of privacy complaints. It should also provide examples or case studies of the way in which it applies those standards in practice.

[91] For more details, see YouTube, Privacy Complaint Process.

[92] See in particular the definition of public interest and Principle 12 of the Global Principles on the Protection of Freedom of Expression and Privacy.

'Fake news'

YouTube does not ban 'fake news' *per se* on its platform. However, YouTube's Community Guidelines make clear that spam, deceptive practices and scams have no place on their platform.[93] In particular, YouTube highlights different types of practices that are banned and may lead to the removal of content under this heading, including: (i) video, channel and comment spam; (ii) artificial traffic spam; (iii) misleading metadata; (iv) misleading or racy thumbnails; (v) scams; (vi) blackmail and extortion.

For instance, YouTube explains that posting large amounts of untargeted, unwanted or repetitive content in videos, comments, private messages or other places on the site is likely to fall foul of its spam policies. This is especially so if the main purpose of the content at issue is to drive people away from YouTube and onto another site.

Similarly, YouTube explains that anything that artificially increases the number of views, likes, comments or other metric, either through the use of automatic systems or by serving up videos to unsuspecting viewers, is against its terms of use. It also bans content that "solely exists to incentivise viewers for engagement (views, likes, comments, etc.)".

In ARTICLE 19's view, YouTube's policies on spam, deceptive practices and scams are broadly consistent with international standards in this area. That YouTube does not attempt to define 'false information' or 'fake news' is to be welcomed. At the same time, we note that YouTube's policies on spam and other deceptive practices appear to be partly designed to address some of the underlying issues driving the dissemination of "false information," such as "clickbait" type content. To that extent, we believe that YouTube should explain more clearly how its policies on spam and "deceptive practices" are related to the broader policy debates on 'fake news' or the dissemination of false information. In particular, the company should be more transparent about the extent to which it might remove 'false information' or "fake accounts" in practice.

We further note that YouTube recently announced that it would seek to promote articles and videos from vetted sources.[94] It is unclear however how YouTube determines which sources are more "authoritative". In any event, YouTube should be more transparent about the algorithm it uses to promote such sources.

Recommendations:

- YouTube should explain more clearly how its policies on spam and "deceptive practices" are related to the broader policy debates on 'fake news' or the dissemination of false information;
- YouTube should be more transparent about the extent to which it might remove 'false information' or "fake accounts" in practice;
- YouTube should explain what it considers to be an "authoritative" source of news and how its algorithm promotes such sources.

Content removal processes: reporting, sanctions and appeals

YouTube has a dedicated webpage outlining enforcement and reporting options.[95] Users can report videos,[96] abusive users,[97] legal complaints,[98] and privacy violations,[99] with other additional reporting tools being available to capture the whole range of content that users may find

[93] See YouTube Community Guidelines, Spam, Deceptive Practices and Scams.
[94] See YouTube Official Blog, *Building a Better News Experience on YouTube, Together*, 9 July 2018
[95] YouTube, Reporting and Enforcement.
[96] YouTube, Report Inappropriate Content.
[97] See YouTube, Reporting and Enforcement.
[98] *Ibid.* More information is available at Removing Content from YouTube.
[99] YouTube Privacy Complaint Process, *op. cit.*

problematic.[100] In other words, different report forms are available depending on the type of complaint at issue (e.g. privacy or legal reporting, reporting on critical injury footage).[101] YouTube, like many other social media companies, also relies on a trusted flagger system, whereby reports filed by trusted flaggers are fast-tracked for review.[102] Finally, YouTube makes clear that it uses "automated detection systems" to flag certain categories of content for removal.[103] This includes child exploitation,[104] 'violent extremism'[105] and copyright material.[106]

Once a potential violation of YouTube's community standards is reported, YouTube reviews the complaint, which may or may not lead to the removal of the video at issue. It appears that YouTube does not generally enable a counter-notice to be filed *before* the material is taken down.

In addition to content being removed, users' account may be penalised by applying community guidelines strikes.[107] In practice, accounts can receive up to three strikes. Each strike entails more limited use of YouTube, including restrictions on live streaming (first strike) and restricted ability to post content for two weeks. Each strikes remains effective for three months. After three strikes within a three-month period, users' accounts are terminated. A separate strike system is in place for copyright violations.[108]

Both Community Guidelines and copyright strike actions can be appealed.[109] Account termination, whether on grounds of community guidelines or copyright can also be appealed.[110] Although users are generally informed that a strike has been applied to their account[111] or that their account has been terminated,[112] it does not appear that they are informed of the specific reasons for the action taken beyond the general explanation in YouTube's policies on community strikes and account terminations.[113]

ARTICLE 19 notes that YouTube's community guidelines enforcement and appeals processes are clearly laid out and easy to find, particularly when compared to other social media platforms such as Facebook or Twitter. However, significant shortfalls remain. In particular, content may be removed before the uploader of the content is given an opportunity to challenge an alleged violation of YouTube's community guidelines. In other words, counter-notices are not possible before action. Moreover, users are not given any specific reasons for actions taken against their content or account, which makes it inevitably harder to appeal YouTube's decisions. It is also noteworthy that the automated character of a three-strike system means that there is less flexibility in ensuring that any sanction is necessary and proportionate in all the circumstances. While some safeguards are built into the system, for instance, the limited lifespan of strikes, it does not entirely sit well with international standards on freedom of expression and due process.

More generally, while YouTube should be commended for providing a snapshot of how it enforces its community guidelines, we believe that the company should provide more information, including disaggregated data about the number of appeals filed and their outcome.

---

[100] YouTube, Other reporting options.
[101] *Ibid.*
[102] YouTube Trusted Flagger Programme.
[103] See YouTube Community Guidelines enforcement report.
[104] *Ibid.*
[105] *Ibid.*
[106] YouTube, How Content ID works.
[107] YouTube, Community Guidelines Strike basics.
[108] YouTube, Copyright Strike basics.
[109] YouTube, Appeal Community Guidelines Actions and Copyright Strike basics, *op. cit.*
[110] YouTube, Account terminations.
[111] See Appeal Community Guidelines Actions, *op. cit.*
[112] See Account terminations, *op. cit.*
[113] *Op. cit.*

Finally, we note that relatively little information is available about the way in which YouTube uses machine learning for the purposes of video flagging. In particular, it is unclear what criteria are used to flag particular pieces of content. This is especially concerning in relation to "violent extremist" content given that automated systems are notoriously bad at understanding context. YouTube should also provide more details about members of its Trusted Flagger Programme.

Recommendations:

- YouTube should ensure that its appeals process complies with the Manila Principles on Intermediary Liability, particularly as regards counter-notices and the giving of reasons for actions taken;
- YouTube should provide disaggregated data on the number of appeals filed and their outcome in its Transparency Report;
- YouTube should be more transparent about its use of algorithms to detect various types of content, such as 'terrorist' videos, 'fake' accounts or 'hate speech';
- YouTube should provide more details about the members of its Trusted Flagger Programme.

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.