Content regulation in the digital age: Submission to the United Nations Special Rapporteur on the right to freedom of opinion and expression

Association for Progressive Communications (APC)

March 2018

Table of contents

1. Introduction

The Association for Progressive Communications (APC) is an international network and non-profit organisation founded in 1990 that works to help ensure everyone has affordable access to a free and open internet to improve lives, realise human rights and create a more just world.

We welcome this topic because it is current and integral to our work. On the one hand there is a lot of "noise" in the mainstream media about so-called "fake news" and what appears to be a fairly rushed response from platforms consisting of increasing in-house regulation of content. On the other hand, human rights defenders and activists we work with express concern that platforms are removing some of their content in a manner that suggests political bias and reinforcing of societal discrimination.

The topic is also particularly important to APC as we continue the process of finding solutions to combating online gender-based violence without such solutions limiting freedom of expression online. Too often, the response to offensive and dangerous though lawful expression is that which seems most simple: censorship, in the form of takedowns, blocking or filtering content. Censorship is increasingly being implemented by private actors, with little transparency or accountability, and disproportionately impacts groups and individuals who face discrimination in society – in other words, groups who look to social media platforms to amplify their voices, form associations, and organise for change. For civil society and multistakeholder forums that deal with content regulation in the digital age more broadly, this is a useful moment to assess the strengths and shortcomings of state regulation and self-regulatory regimes when it comes to protecting the wide range of rights that internet users around the world have come to rely on to exercise their rights online and offline.

2. Company compliance with state laws

As the internet becomes increasingly ubiquitous it is not surprising that it is being used to deliberately spread misinformation, disrupt electoral processes, or recruit terrorists. It is also no surprise that internet platforms are facing unprecedented pressure to comply with state laws to regulate content. In fact, online platforms are subject to opposing demands: "one asking them to thoroughly police the content posted on their services to guarantee the respect of national laws, and the other objecting to them making determinations on their own and exercising proactive content monitoring, for fear of detrimental human rights implications. Moreover, given that the current non-liability regimes were initially established for 'passive' intermediaries, the fear of a potential loss of protection may disincentivise companies from assuming more responsibilities."[1]

APC underscores that companies have the responsibility to respect human rights. This means they should refrain from infringing on human rights and take measures to address adverse human rights impacts resulting from their business models, policies, practices, and the services they provide.[2] Companies should, therefore, not comply with measures imposed by states that are not consistent with Article 19 of the International Covenant on Civil and Political Rights (ICCPR). However, we observe that under increasing pressure, companies are not only complying with state laws concerning content regulation and

[1] Internet and Jurisdiction Policy Network. (2017). *Content and jurisdiction policy options: Cross-border content restrictions*. https://www.internetjurisdiction.net/uploads/pdfs/Papers/Content-Jurisdiction-Policy-Options-Document.pdf

[2] Ruggie, J. (2011). *UN Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework*. www.ohchr.org/EN/Issues/TransnationalCorporations/Pages/Reports.aspx

other measures imposed by governments; they appear to also take pre-emptive measures through, for example, adaptations to their terms of service agreements.

2.1. Terrorism-related and extremist content

Companies are removing content deemed to be terrorism-related and extremist to comply with state laws, which raises concern both because laws themselves are overbroad and restrict permissible speech, and because laws impose large penalties and unrealistically quick responses to unlawful speech, which encourages companies to remove questionable content in order to avoid being fined. An example of the former is Pakistan's Prevention of Electronic Crimes Act (PECA). PECA provides the Pakistan Telecommunications Authority (PTA) with complete authority (without independent or judicial oversight) to directly block whatever it considers to be "objectionable content".[3] According to Article 9, content that glorifies an offence or a person accused of a crime, or that supports terrorism or activities of a terrorist organisation, is prohibited. Section 12 criminalises preparation or dissemination of "information, through any information system or device that invites or motivates to fund, or recruits people for terrorism or plans for terrorism." The ambiguity of language means that compliance with PECA could require platforms to remove content that merely disseminates information about a terrorist incident or discusses newsworthy events, i.e. permissible expression under international human rights norms. Section 37 of PECA, which does not criminalise content per se, defines very broad categories of "unlawful content" that is supposed to be proactively blocked by the PTA.[4]

An example of the latter concern is the new German "network enforcement law", or NetzDG, which went into full effect in January 2018 and requires internet platforms with more than two million users to delete threats of violence and slander within 24 hours of a complaint being received, or within seven days if cases are more legally complex.[5] Fines for non-compliance amount to up to €50 million (USD 60 million). Such short time frames combined with steep penalties means that platforms will err on the side of caution, removing lawful content rather than be found in non-compliance. Furthermore, under the new law, decision making on what content should be removed, which previously fell within the role and responsibility of the judiciary, has now been delegated to private platforms. This removes judicial oversight and leaves users with little recourse to challenge removals, since with the new system, it is not clear why content is being removed. Since NetzDG was introduced, other governments, including Russia and the United Kingdom, have indicated interest in pursuing similar approaches.[6]

At the European Union level, Facebook, Twitter, Microsoft and YouTube agreed in 2016 to a new Code of Conduct that requires them to review "the majority of" hateful online content within 24 hours of being notified, and to remove it if necessary in the name of combating hate speech and terrorist propaganda across the EU. The Code of Conduct puts more responsibility on platforms to police content, without the

[3] The International Center for Not-for-Profit Law. (2015). *Analysis of Pakistan's Prevention of Electronic Crimes Bill, 2015*. www.icnl.org/research/library/files/Pakistan/pecacomms.pdf

[4] Khan, S. (2018). Legal limitations on online expression in Pakistan. In APC, *Unshackling Expression: A study on laws criminalising expression online in Asia*. https://www.giswatch.org/2017-special-report-unshackling-expression-study-law-criminalising-expression-online-asia

[5] Deutsche Welle. (2018, 1 January). Germany implements new internet hate speech crackdown. www.dw.com/en/germany-implements-new-internet-hate-speech-crackdown/a-41991590

[6] Reporters Without Borders. (2017, 19 July). Russian bill is copy-and-paste of Germany's hate speech law. https://rsf.org/en/news/russian-bill-copy-and-paste-germanys-hate-speech-law

accountability and oversight of democratic institutions, or safeguards to ensure that lawful content (for example, journalism covering the topic of extremism) is not arbitrarily taken down.[7]

In addition to state regulations, companies are entering into self-regulatory mechanisms, like the Global Internet Forum to Counter Terrorism, through which Facebook, Microsoft, YouTube and Twitter are collaborating to curtail the spread of terrorism and violent extremism through technical solutions, research, knowledge sharing,[8] and building on the Shared Industry Hash Database.[9]

2.2. False news, disinformation and propaganda

APC is among the growing voices of civil society concerned about the direction of international discussions on "fake news" and the possible paths that the framing of the issue is taking.[10] This is especially critical around elections, when the impact of the spread of false news, disinformation and propaganda on democratic institutions can be significant and harmful. We recognise that as states seek to address this challenge, platforms may face increasing pressure to take down content that is legitimate political expression.[11]

Some platforms have also taken measures to counter so-called "fake news". For example, Facebook announced in January 2018 that it planned to prioritise high-quality news on the social network by allowing its users to rank news sources that they see as the most credible and trustworthy.[12] This move is in response to what Mark Zuckerberg described as "too much sensationalism, misinformation and polarization in the world today".[13] As the Special Rapporteur pointed out, this raises serious concerns in various parts of the world where news sources might be deemed trustworthy by a particular community but those sources are censored or illegal in that country.[14] Even where independent news is not outlawed, Facebook's new ranking system may be open to manipulation, and could result in lesser-known outlets and alternatives to mainstream news outlets being buried to the point of obscurity in Facebook's feed. The ranking system comes on the heels of an announcement by Facebook of plans to remove posts made

[7] Toor, A. (2016, 31 May). Facebook, Twitter, Google, and Microsoft agree to EU hate speech rules. *The Verge*. https://www.theverge.com/2016/5/31/11817540/facebook-twitter-google-microsoft-hate-speech-europe

[8] Twitter Public Policy. (2017, 26 June). Global Internet Forum to Counter Terrorism. https://blog.twitter.com/official/en_us/topics/company/2017/Global-Internet-Forum-to-Counter-Terrorism.html

[9] In December 2016, Facebook, Microsoft, Twitter and YouTube created a shared industry database of "hashes" – unique digital "fingerprints" – for violent terrorist imagery and terrorist recruitment videos or images in order to help identify potential terrorist content on their respective platforms. https://blog.google/topics/google-europe/partnering-help-curb-spread-terrorist-content-online

[10] Coding Rights, et. al. (2017). Open Letter from Latin American and Caribbean Civil Society Representatives on the Concerns around the Discourse about Fake News and Elections. https://direitosnarede.org.br/c/openletter-latinamericacivilsociety-ifg2017

[11] In Kenya, for example, according to the Kenya ICT Action Network, the Communications Authority in conjunction with the National Cohesion and Integration Commission published Guidelines on Prevention of Dissemination of Undesirable Bulk and Premium Rate Political Messages and Political Social Media Content Via Electronic Communications Networks that require takedown of political messages with "undesirable content". The definition of undesirable content is borrowed from one of the Authority's licence conditions, which apply to the whole spectrum of bad information/misinformation (unintentional false content), disinformation (false content intended to harm) and malinformation (factual content intended to harm). Lumping together different subjects without consideration of the intent and consequences endangers freedom of expression for legitimate speech such as artistic and academic content. See: Kenya ICT Action Network. (2018). *Moving forward while looking back: Freedom online in Kenya's 2017 elections*. https://www.kictanet.or.ke/?sdm_downloads=moving-forward-while-looking-back

[12] Frenkel, S., & Maheshwari, S. (2018, 19 January). Facebook to Let Users Rank Credibility of News. *The New York Times*. https://mobile.nytimes.com/2018/01/19/technology/facebook-news-feed.html?login=email&auth=login-email

[13] Zuckerberg, M. (2018, 19 January). Continuing our focus for 2018 to make sure the time we all spend on Facebook is time well spent... https://www.facebook.com/zuck/posts/10104445245963251?pnref=story

[14] Frenkel, S., & Maheshwari, S. (2018, 19 January). Op. cit.

by “pages”, including those of independent news organisations, from users' regular news feed. This move was ostensibly not in response to the challenge of disinformation and propaganda flooding the social network, but rather to maximise “meaningful interactions” users have on the platform by prioritising posts from friends and family; however, Facebook’s experiment with this new system in Serbia already resulted in an existential threat to independent media.[15]

Google search result algorithms have also been implicated in apparent bias through how news sources are ranked in response to a user request. A 2016 investigation by *The Guardian* found that “Google’s search algorithm appears to be systematically promoting information that is either false or slanted with an extreme right wing bias on subjects as varied as climate change and homosexuality.”[16] Similar observations have been made about Google’s autocomplete function.[17] This is not intentional, and Google does try to fix specific instances when they are brought to their attention. However, as the authors of the *Guardian* article point out, these fixes are made quietly by humans at Google through “manual adjustments in a process that’s neither transparent nor accountable.” The *Guardian* investigators also point out that politically motivated third parties, including the “alt-right” movement in the United States, “use a variety of techniques to trick the algorithm and push propaganda and misinformation higher up Google’s search rankings.”

Google launched an effort called “Project Owl” in an attempt to address this problem in April 2016.[18] Project Owl is an automated effort to eliminate so-called fake news sources from search results and elevate authoritative news sources in result rankings. It has been reported that alternative news sources like the left-wing “World Socialist Web Site” have experienced dramatic reductions in their internet traffic as a result.

It remains to be seen what the impact of these measures will be over time. Irrespective of their effectiveness, they should be seen in the context of the broader business model of these platforms. Facebook is a social network that evolved organically into also being a news distribution network. A platform which qualified for protection from liability as an intermediary also became the world’s biggest-ever content distributor.

In both the Google and Facebook examples mentioned above, the problems they are trying to address were also created by them in the first place. While users are the sources of the news content that is distributed on the platform in the case of Facebook, and in the case of Google it is “just” a search engine, the reality is that both platforms – Facebook perhaps more so than Google – do interact with users and content through, for example, the use of algorithms to shape the news feeds or search results that users receive and by directing specific advertising content to users. They are not the passive intermediaries that current intermediary liability regimes were constructed for, nor are they publishers of content in the way that a newspaper is. We are not suggesting that the response to this dilemma should be to make

[15] Dojcinovic, S. (2017). Hey, Mark Zuckerberg: My Democracy Isn’t Your Laboratory. *The New York Times*. https://mobile.nytimes.com/2017/11/15/opinion/serbia-facebook-explore-feed.html?smid=tw-share&referer=https://t.co/eFScPDzO2T?amp=1

[16] Solon, O., & Levin, S. (2016, 16 December). How Google's search algorithm spreads false information with a rightwing bias. *The Guardian*. https://www.theguardian.com/technology/2016/dec/16/google-autocomplete-rightwing-bias-algorithm-political-propaganda

[17] Lapowsky, I. (2018, 12 February). Google Autocomplete Still Makes Vile Suggestions. *Wired*. https://www.wired.com/story/google-autocomplete-vile-suggestions/amp?__twitter_impression=true

[18] Wakabayashi, D. (2017, 26 September). As Google Fights Fake News, Voices on the Margins Raise Alarm. *The New York Times*. https://www.nytimes.com/2017/09/26/technology/google-search-bias-claims.html

platforms such as Facebook legally liable for the content carried on the platform, but there is a clear need for more transparency and accountability in how they manage and manipulate content and user data.

Also relevant to the topic of false news, disinformation and propaganda is the role that third party actors such as data brokers, market research firms and advertising agencies play in the processing, use and management of content, users and user data. Disinformation campaigns use the same targeted internet advertising system, and probably the same data brokers, used by familiar brands. These data brokers gather personal data such as past purchases, petitions signed, sites visited and news sources clicked on from multiple sources and across devices, and there is currently, outside of Europe, very little regulation of how they operate and not much scrutiny of whether they comply with human rights. Former US ambassador to the Organisation for Economic Co-operation and Development (OECD) Karen Kornbluh proposes that the new EU General Data Protection Regulation (GDPR) could be utilised to curb online disinformation by helping "limit the potency of disinformation without the need for a judge or platform to adjudicate what is or is not hate speech or fake news."[19] This might simply boil down to making the role these third party actors play more transparent, or it could involve regulatory intervention. Her suggestion is grounded in provisions in the GDPR "that could restrict how bad actors tailor disinformation by making it harder for them to use the personal data they need to make their campaigns effective in targeting susceptible individuals."

2.3. The "right to be forgotten" framework?

APC recognises that the request to be de-listed,[20] when rooted in data protection frameworks that provide safeguards for freedom of expression and access to information, has benefits for the right to privacy and autonomy over data.[21] We also recognise, however, that the request to be de-listed risks abuse and threatens freedom of expression and access to information when it is applied without such safeguards, and when different criteria are applied (i.e. lacking a provision excluding information in the public interest, or pertaining to public figures). For example, the request to be de-listed is being considered in jurisdictions, like India, where data protection frameworks are not in place.[22] This is problematic as data protection frameworks should provide for procedural safeguards and limitations that protect against the de-listing of information in the public interest. In Russia, for example, legislation was passed in the aftermath of the Google v. Spain case, which requires the deletion of posts from their original websites, not just de-listing them from search engines. It represents a significant threat to freedom of expression and access to information, as it not does exclude information related to a public figure or in the public interest.[23]

[19] Kornbluh, K. (2018, 20 February). Could Europe's New Data Protection Regulation Curb Online Disinformation? *Council on Foreign Relations*. https://www.cfr.org/blog/could-europes-new-data-protection-regulation-curb-online-disinformation

[20] APC considers the term the "right to be forgotten" a misnomer and misleading. We understand it to mean the request to be de-listed, which is a remedy that enables individuals to request to be de-listed from search results produced on the basis of a search term which includes their name. This remedy has been derived from the right to erasure under data protection law by some international and domestic courts.

[21] For more specific factors we would like to see considered, see Principle 18 of ARTICLE 19's Global Principles on Protection of Freedom of Expression and Privacy: article19.shorthand.com

[22] Sinha, A. (2017, 7 April). Right to be Forgotten: A Tale of Two Judgements. *Centre for Internet & Society*. https://cis-india.org/internet-governance/blog/right-to-be-forgotten-a-tale-of-two-judgments

[23] Deutsche Welle. (2015, 7 March). Russian parliament approves 'right to be forgotten online' law. www.dw.com/en/russian-parliament-approves-right-to-be-forgotten-online-law/a-18560565

As we note in section 4, APC does not support the global application of de-listing requests. No single government should be able to decide what people in the rest of the world can see in their search results. The global application of de-listing requests would amount to the removal of content from search results on the basis of a law in a particular jurisdiction, rather than on the basis of limitations to freedom of expression rooted in international human rights law. Imposing the request to de-list on jurisdictions that have not recognised it undermines domestic protections for human rights, especially in countries where it can be used by political actors to erase inconvenient aspects of national history.

2.4. How should companies respond to state content regulation laws and measures that may be inconsistent with international human rights standards?

Companies should not comply with requests that are inconsistent with international human rights standards, and that undermine due process. Compliance with state laws needs to be seen in the context of many of these laws being new, often quite vague, and/or applied in a very arbitrary manner. Platforms' readiness to comply with state regulation runs the risk of legitimising such laws and reinforcing the trend for states to regulate online speech in the first place.

We recognise that by failing to comply with requests for content regulation from states, companies face having to withdraw offering their services in specific jurisdictions and that their employees face threats of arrest. Therefore we recommend the following considerations/steps when responding to such requests:

- Evaluating whether regulating content is actually necessary to comply with national law.
- Ensuring that proper procedures were followed and a judicial order was issued, and pushing back against requests where procedures were not followed.
- Ensuring that any regulation of content constitutes the least restrictive measure (i.e. blocking one tweet rather than an entire account, use of geoblocking, etc.).
- Ensuring that complying with national law does not put users at risk of imprisonment or other forms of persecution.
- When evaluating whether to operate in the jurisdiction, companies should carry out human rights impact assessments (HRIAs) to mitigate risks, and to determine whether complying with state content regulation laws does more to facilitate or restrict freedom of expression. HRIAs should not be a one-time occurrence, but should be carried out on a periodic basis to re-assess risks, measures needed to mitigate risks, and the feasibility of meeting their responsibility to respect human rights while operating in the jurisdiction, especially as changes in national law occur.
- Being transparent about regulation of content: reporting on takedown requests and other forms of content restrictions in the form of transparency reports; being transparent about any agreements entered into with states; notifying users that content has been restricted and why.

3. Other state requests

3.1. State requests based on platform terms of service (ToS) and "shadow" requests

In addition to content removals on the basis of national law, states request the removal of content on the basis of terms of service (ToS) requests, including through government-aligned actors, as well as through secretive agreements between the state and platforms.

The line is often blurred between ToS requests coming from private actors and those coming from the government, since company reporting in this area is lacking.[24] Also, according to the 2017 Ranking Digital Rights Corporate Accountability Index, companies tended to disclose more information about requests they receive from governments and private parties to restrict or delete content or deactivate accounts than about actions companies themselves took to enforce terms of service. However, state actors may also use ToS requests, with little transparency.[25] Disclosure about private requests for content restriction is important for monitoring the full impact of government requests for content restriction, given that governments often delegate takedown requests and the reporting of ToS violations to private actors. For example, there have been documented cases of copyright enforcement mechanisms being abused by governments, such as Ecuador's President Rafael Correa, who used millions of dollars of public funds to hire a foreign company to help delete information critical of him from sites including YouTube, Facebook, Vimeo and Dailymotion. In addition, in May 2015, there were allegations from thousands of Ukrainians that Russian trolls had misused Facebook's reporting mechanism to take down content from Ukraine during the Russia-Ukraine conflict. Facebook maintained that it did the right thing according to its policies in taking down the posts, citing hate speech provisions.[26]

According to the Global Network Initiative (GNI), ToS enforcement decisions by GNI member companies[27] do not change based on whether the allegedly inappropriate content is referred to the companies by governments or by any other third party.[28] Thomas Myrup Kristensen, head of Facebook's office in Brussels, denied that large numbers of complaints influence its decisions: "It doesn't matter if something is reported once or 100 times, we only remove content that goes against [our] standards."[29] However, it is APC's experience that reports from "ordinary" users are not given the same weight. Abuse, particularly non-consensual image sharing, is rampant, and reports are often rejected with an explanation that the abuse did not violate Facebook's community guidelines even though leaked documents show they are clear violations.[30]

State requests come in other forms of "shadow regulation". For example, under pressure from the UK Intellectual Property Office, search engines agreed in 2017 to a "Voluntary Code of Practice" that requires them to take additional steps to remove links to allegedly unlawful content. Domain name registrars are placed under pressure to participate in copyright enforcement, including by "voluntarily" suspending domain names.[31]

[24] The 2017 Ranking Digital Rights report found that company disclosure is inadequate across the board, including disclosure related to ToS-related removal requests. See: https://rankingdigitalrights.org/index2017/findings/keyfindings

[25] Ibid.

[26] Hovyadinov, S. (2018, 25 January). When Transparency Also Needs Transparency. *New America Weekly*. https://www.newamerica.org/weekly/edition-191/when-transparency-also-needs-transparency

[27] GNI member companies include Facebook, Google, Microsoft, and Oath, among others relevant for this study. See: https://globalnetworkinitiative.org/participants/index.php?qt-gni_participants=1#qt-gni_participants

[28] Global Network Initiative. (2016). *Extremist Content and the ICT Sector: A Global Network Initiative Policy Brief*. https://globalnetworkinitiative.org/sites/default/files/Extremist-Content-and-the-ICT-Sector.pdf

[29] Hovyadinov, S. (2018, 25 January). Op. cit.

[30] Association for Progressive Communications. (2017). Statement on Facebook's internal guidelines for content moderation. https://www.apc.org/en/pubs/statement-facebooks-internal-guidelines-content-moderation

[31] McSherry, C., York, J. C., & Cohn, C. (2018, 30 January). Private Censorship Is Not the Best Way to Fight Hate or Defend Democracy: Here Are Some Better Ideas. *Electronic Frontier Foundation*. https://www.eff.org/deeplinks/2018/01/private-censorship-not-best-way-fight-hate-or-defend-democracy-here-are-some

APC agrees with the Electronic Frontier Foundation (EFF) that shadow regulation is dangerous and undemocratic. We support their recommendations that regulation should take place in the open, with the participation of the various interests that will have to live with the result. "To help alleviate the problem, negotiators should seek to include meaningful representation from all groups with a significant interest in the agreement; balanced and transparent deliberative processes; and mechanisms of accountability such as independent reviews, audits, and elections."[32]

3.2. Non-transparent agreements with companies

States enter into secretive agreements with companies on a range of issues, from copyright to blasphemy to violent extremism, to voluntarily remove content from their platforms. Such measures bypass critical democratic institutions, like the judiciary; have the potential to censor legitimate speech, including journalistic reporting on matters of public interest and counter speech; and impose liability on intermediaries, which can lead to over-compliance and removal of permissible speech for fear of penalties. In addition, there is the question of the effectiveness of such measures.

According to the GNI Implementation Guidelines, GNI member companies are expected to refrain from entering into voluntary agreements that require the participants to limit users' freedom of expression or privacy in a manner inconsistent with its Principles on Freedom of Expression and Privacy. Voluntary agreements entered into prior to committing to the Principles and which meet this criterion should be revoked within three years of committing to the Principles.[33] However, there are reports that such agreements remain in place.

For example, In September 2016, Israeli Public Justice Minister Ayelet Shaked announced that close cooperation between the Israeli government and Facebook would take place to tackle "incitement" online. This involved encouraging social media networks to remove all content that Israel deems "incitement". The term incitement has been vaguely defined by Israel, but can include discourse and rhetoric that resists or criticises Israeli policy. The cyber unit at Israel's State Attorney's Office reported that in 2017, 85% of the Israeli government's requests to remove content were accepted, representing an increase from 70% in 2016.[34] The cyber unit works in close cooperation with both Facebook and Twitter to censor and remove online content that is perceived as "inciteful". According to Adalah – The Legal Center for Arab Minority Rights, the unit removed 1,554 cases of online content in 2016, constituting a grave violation of Israeli Basic Law which states, "Nothing in the law allows state authorities to censor content based solely on an administrative determination." This censorship undertaken by the state therefore amounts to an illegal offence.[35]

4. Global removals

APC considers global removals and de-listings to be an exceptional measure, which should only be applied when the content is in violation of international human rights standards and removing the content

[32] Ibid.

[33] See Global Network Initiative Implementation Guidelines: https://www.globalnetworkinitiative.org/sites/default/files/Implementation-Guidelines-for-the-GNI-Principles_0.pdf

[34] Ilan, S. (2017, 29 December). Israel Official Reports Increased Cooperation on Removing Content from Social Media. *Ctech*. https://www.calcalistech.com/ctech/articles/0,7340,L-3728439,00.html

[35] Adalah. (2017, 14 September). Israel's 'Cyber Unit' operating illegally to censor social media content. https://www.adalah.org/en/content/view/9228

globally is a necessary and proportionate response to prevent or mitigate the harm it would inflict. For example, when the non-consensual dissemination of intimate content constitutes a violation of a person's right to be free from violence, the content should be removed (or if not possible, de-listed) at a global level, as is the policy of platforms like Google.[36]

Companies should not be removing access to content unless it violates international human rights law. Understanding that the permissibility of content can be contextual (meaning content might rise to the level of incitement to hatred, discrimination or violence in one jurisdiction, but not in another) and that companies may need to comply with national law, content removals should take on the least restrictive form, limited to only the jurisdiction where the removal is legally necessary to protect any individual rights. When assessing the proportionality of a restriction to freedom of expression on the internet, the impact that the restriction could have on the internet's capacity to guarantee and promote freedom of expression must be weighed against the benefits that the restriction would have in protecting other interests.[37] Global removals of content generally fail to meet the strict proportionality test.[38]

5. Individuals at risk

APC does not believe that company standards adequately reflect the interests of users at risk. Here are some examples:

5.1. Protected categories

According to leaked internal documents published in May 2017 by *The Guardian*, which revealed how Facebook moderates content, Facebook's protected categories do not adequately protect users at risk.[39] According to the leaked guidelines, "protected categories" are defined based on race, sex, gender identity, religious affiliation, national origin, ethnicity, sexual orientation and serious disability/disease. First, these categories are not fully consistent with international human rights law. For example, caste is not considered a protected category, even though the Committee on the Elimination of Racial Discrimination (CERD) considers it to be.[40] At-risk groups like migrants, refugees and asylum seekers are regarded as a "quasi-protected category",[41] so they do not receive the protections given to other

[36] According to the Cyber Civil Rights Initiative, Facebook, Google, Instagram, Reddit, Tumblr, Twitter and Yahoo do not allow non-consensual porn on their platforms, though we also note how frequently user reports of non-consensual images do not receive a satisfactory response from these same platforms. See the Cyber Civil Rights Initiative Online Removal Guide: https://www.cybercivilrights.org/online-removal

[37] Special Rapporteur on Freedom of Opinion and Expression, Organization for Security and Co-operation in Europe (OSCE) Representative on Freedom of the Media, Organization of American States (OAS) Special Rapporteur on Freedom of Expression, & African Commission on Human and Peoples' Rights (ACHPR) Special Rapporteur on Freedom of Expression and Access to Information. (2011). Joint Declaration on Freedom of Expression and the Internet. https://www.oas.org/en/iachr/expression/showarticle.asp?artID=848

[38] For further analysis, see: Internet Freedom Foundation, et al. (2017). Joint legal submission before the French Council of State (Conseil d'État). https://www.apc.org/sites/default/files/Google_France_Intervention_English.pdf

[39] The Guardian. (2017). Facebook Files. https://www.theguardian.com/news/series/facebook-files

[40] Section 4 of the Committee on the Elimination of Racial Discrimination's General Recommendation 29 says that states must "(r) Take measures against any dissemination of ideas of caste superiority and inferiority or which attempt to justify violence, hatred or discrimination against descent-based communities" and "(s) Take strict measures against any incitement to discrimination or violence against the communities, including through the Internet." See: Committee on the Elimination of Racial Discrimination. (2002). General recommendation XXIX on article 1, paragraph 1, of the Convention (Descent). tbinternet.ohchr.org/_layouts/treatybodyexternal/Download.aspx?symbolno=INT%2fCERD%2fGEC%2f7501&Lang=en

[41] Section 3 of the Committee on the Elimination of Racial Discrimination's General Recommendation 30 compels states to "11. Take steps to address xenophobic attitudes and behaviour towards non-citizens, in particular hate

vulnerable groups – however, “heterosexuals” are a protected category. Second, the way in which Facebook treats “subsets” of groups is also concerning. If a protected category (Muslim person) is combined with an unprotected category (child) or quasi-protected category (refugee), then the lower level of protection (i.e. none, or quasi) is afforded to them, rather than the maximum. To put it another way, people who face multiple forms of discrimination are penalised for it. Third, Facebook’s approach of defending all races, genders and other protected categories equally overlooks the complexity of power dynamics in society, which may leave certain races or genders more at risk, and disregards the multiple forms of discrimination that may be playing out. This may result in further perpetuation of discrimination.[42]

5.2. Online gender-based violence

As noted previously, abuse, particularly non-consensual image sharing, is rampant online, and reports are often rejected with an explanation that the abuse did not violate community guidelines or ToS. For many users, this is tantamount to being told that the abuse experienced did not take place. This appears to be a result of how community guidelines are understood and implemented. For example, Facebook's “revenge porn” guidelines do not reflect an understanding of harm in different contexts.[43] We know from experience that human rights defenders are frequently silenced by Facebook itself and face a wide variety of abuse from fellow users, such as the creation of imposter profiles that discredit or defame, photo alteration to create fake intimate images, hate speech, threats and doxxing. When Facebook requires that the image involve sexual activity, it does not seem to consider that the definition of such activity can be different in different communities. Images that may be perfectly acceptable in one community may constitute risk for a woman in another community. The platform fails to recognise that what is most important is whether or not the person in the image finds it to be non-consensual and faces the risk of harm.[44] Additionally, in part because reporting is isolated and de-contextualised, the use of “credible” harm as a measure of whether content should be taken down is problematic. Reporting systems tend to be focused on individual posts that are de-contextualised, and there is insufficient information in one report to understand if a threat is credible or not. Furthermore, reporting systems do not take into account an understanding that rape culture and a culture of femicide make the risk posed by such content much more credible to users than to the platform, and a lack of awareness of report review staff in this area will lead them to assume that most threats are not credible.

speech and racial violence, and to promote a better understanding of the principle of non-discrimination in respect of the situation of non-citizens” and “Take resolute action to counter any tendency to target, stigmatize, stereotype or profile, on the basis of race, colour, descent, and national or ethnic origin, members of 'non-citizen' population groups, especially by politicians, officials, educators and the media, on the Internet and other electronic communications networks and in society at large.” See: Committee on the Elimination of Racial Discrimination. (2005). General recommendation XXX on discrimination against non-citizens. tbinternet.ohchr.org/_layouts/treatybodyexternal/Download.aspx?symbolno=INT%2fCERD%2fGEC%2f7502&Lang=en

[42] For more examples and analysis, see: Angwin, J., & Grassegger, H. (2017, 28 June). Facebook’s Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children. *ProPublica*. https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms

[43] The Guardian. (2017, 22 May). What Facebook says on 'sextortion' and 'revenge porn'. https://www.theguardian.com/news/gallery/2017/may/22/what-facebook-says-on-sextortion-and-revenge-porn

[44] Association for Progressive Communications. (2017). Statement on Facebook's internal guidelines for content moderation. https://www.apc.org/en/pubs/statement-facebooks-internal-guidelines-content-moderation

5.3. LGBTIQ people and sexual rights activists

In 2017, APC published our third EROTICS[45] Global Survey on Sexuality, Rights and Internet Regulations.[46] The survey reached out to respondents who broadly self-identified as "working on" LGBTIQ, women's and sexual rights, which potentially included activists, scholars, experts and supporters. The final quantitative sample included 332 respondents.[47] Among other findings, 66% of the sample said that the internet is considered an "important" or "very important" medium of sexual expression. The top purposes cited for using the internet were to share critical information quickly and widely (84%), and to search for information that is difficult to find in offline spaces (82%). Top services used by respondents were social networks (98%), followed by instant messaging and email (both 92%), other websites (89%), blogs (87%) and hosting services (85%). The majority of respondents reported that they had experienced threats online, such as harassment (75%), intimidating online comments (63%) and blocked websites or filtering software that prevented the user from accessing information (54%).[48] Of the 76 responses regarding where threats were experienced, social networks were most commonly cited, with Facebook in first place (59%), followed by Twitter (16%). Respondents cited the following topics as most frequently regulated, censored and/or monitored: pedophilia (81%), anti-government, abortion (both 68%) and "obscene" content (67%). The most common reasons given by the government or/and corporations to regulate, prohibit, remove and/or censor content that respondents search for, share or produce on the internet included public decency (52%), followed at a long distance by anti-terrorism (27%) and preserving tradition (22%). Respondents overall do not consider the internet a safe place and consider that corporations do very little or nothing when they receive complaints of threats from their users. Nonetheless, 88% of them consider that the internet enables and increases the power, visibility, communication and organisation of women and minorities.

5.4. Religious minorities and non-conformists

The right to freedom of expression of religious minorities, those belonging to majority communities holding liberal views, secularists and atheists is restricted on platforms due to religious sensitivities, opposition from a large number of individuals belonging to the majority (through flagging of content for takedown, or a flood of abusive responses), and fear of state regulations. One of the most common bases for these takedown requests relates to content deemed objectionable from a religious point of view. Blasphemy laws inherently violate freedom of expression and are the most commonly cited instrument for platforms to "proactively" take down content. Beyond blasphemy laws, states and non-state actors have been relying on other means to shut down expression in this regard. For example, Google has entered into an agreement with the government of Pakistan to remove blasphemous content for a local version of YouTube in order to lift the three-year ban on access to the platform in the country.[49] Such

[45] EROTICS is a global network of activists, academics and organisations working on sexuality issues including LGBTIQ rights, sex work and sex education, among others. See: https://www.apc.org/en/project/erotics-exploratory-research-project-sexuality-and-internet

[46] Association for Progressive Communications. (2017). *EROTICS Global Survey 2017: Sexuality, rights and internet regulations*. https://www.apc.org/sites/default/files/Erotics_2_FIND-2.pdf

[47] The majority of the sample – 40% – lives in Latin America and the Caribbean (LAC); 21% in South, South East and East Asia (SA); 20% in Africa; 12% in North America and Western Europe; 4% in Western Asia (WA) and 2% in Eastern Europe.

[48] Percentages represent the combined responses of "sometimes", "often" and "always".

[49] Wilkes, T. (2016, 18 January). Pakistan lifts ban on YouTube after launch of local version. *Reuters*. https://www.reuters.com/article/us-pakistan-youtube/pakistan-lifts-ban-on-youtube-after-launch-of-local-version-

local agreements allow states to exercise greater control over content and remove the possibility for redress.[50] As noted previously, such agreements are not subject to safeguards, accountability, and the right to remedy, and risk censoring legitimate speech that is critical or controversial.

5.5. Silencing discussion of hate speech

Minority or marginalised groups use platforms to call out racism, start a dialogue, and reclaim derogatory language that has been used to target them. Facebook routinely takes down such content, and also places users in "Facebook jail", locking them out of their accounts for 24 hours or longer. Facebook acknowledged this problem with the "Hard Questions: Hate Speech" blog post published on 27 June 2017,[51] but news reports in the US indicate that the problem is continuing and far-reaching.[52] Outside the US, where context is even more difficult for content moderators to comprehend due to linguistic and cultural nuances, we anticipate this is even more difficult to get right.

5.6. Language and cultural context

Companies struggle to keep up with slang and harassment trends, support all user languages and understand different cultural contexts. Speakers of minority languages may face greater harm related to online abuse because their reports are rejected or the reporting mechanism is not in their language.

6. Content regulation processes

In APC's view, users are not given an explanation (or a sufficient explanation) as to why their content is taken down, which means they are not able to avoid such takedowns in the future. Implementation of content restriction and takedowns is still mostly not clear, accessible or easily understandable and is lacking in accountability and due process. Greater transparency would be an important step to empowering users to challenge the removal of their content and access remedy. In its 2017 index, Ranking Digital Rights found that only three companies disclosed any data about the volume and nature of content they removed at their own initiative when enforcing their terms of service – Google, Twitter and Microsoft.

7. Bias and non-discrimination

Offline power structures are often replicated online, and are reflected in how ToS are implemented and how users are able to access remedy. For example, we observe that in community standards, concepts of nudity are influenced by social and cultural norms that are gender biased. This is not just Facebook's censorship of women's nipples, but also women showing body hair or even bodily functions such as menstruation are more likely to be censored. Women's nudity is automatically sexualised, women's bodies are objectified, and nudity used for political expression and women's agency is included in this "sexual content" category.[53]

idUSKCN0UW1ER

[50] Association for Progressive Communications. (2018). *State of the Internet in Asia: The case of India, Malaysia and Pakistan*. https://www.apc.org/en/pubs/state-internet-asia-case-india-malaysia-and-pakistan

[51] Allen, R. (2017, 27 June). Hard Questions: Who Should Decide What Is Hate Speech in an Online Global Community? *Facebook Newsroom*. https://newsroom.fb.com/news/2017/06/hard-questions-hate-speech

[52] See, for example: Oluo, I. (2017, 2 August). Facebook's Complicity in the Silencing of Black Women. *Medium*. https://medium.com/@IjeomaOluo/facebooks-complicity-in-the-silencing-of-black-women-e60c34434181

The way in which Facebook deals with hateful and inciteful content in Israel and Palestine demonstrates how biases influence content moderation. Facebook's stated policy is to "remove content, disable accounts and work with law enforcement when [they] believe that there is a genuine risk of physical harm or direct threats to public safety"[54] and that content is removed if it attacks people based on their religion, race, ethnicity or national origin. Research conducted by the Arab Center for Social Media Advancement – 7amleh, a Palestinian NGO and APC member, revealed that the number of hateful or inciteful posts uploaded by Israelis on social media doubled from 280,000 in 2015 to 675,000 in 2016. The majority of posts contained the words killing, followed by death and expulsion. A hateful post is uploaded by Israelis every 46 seconds, and the majority of them are posted on Facebook.[55] Nevertheless, despite heightened Israeli incitement online towards Arab and Palestinian communities, the number of removed posts, temporarily or permanently suspended accounts or deleted pages was zero, amounting to a flagrant violation of Facebook community standards.[56] This may be attributed to the fact that Facebook has granted Zionists the status of a "globally protected group", according to a *Guardian* article on Facebook's manual of credible threats of violence.[57] In other words, Israeli citizens are fully protected by Facebook policies when publishing content online, while Palestinians, a people to whom basic human rights have already been denied, are subject to close monitoring and account removal.

As noted in section 5, Facebook's approach of defending all protected categories equally, without contextualisation, may result in further perpetuation of discrimination. This bias prompted a joint letter by civil society organisations to Facebook CEO Mark Zuckerberg protesting the policy of removing content of marginalised communities at the request of government actors,[58] as Facebook's collaboration with government actors further exacerbates the already existing power inequality between government actors and marginalised groups.

8. Appeals and remedies

Users are not consistently being notified about content restrictions, takedowns and account suspensions, the reasons for these actions, and the procedures they must follow to seek reversal of such actions. APC has examples from research on online gender-based violence of women not being notified about the decision regarding their requests to have content removed.[59]

[53] See, for example: Chemaly, S. (2017, 6 December). Why Female Nudity Isn't Obscene, But Is Threatening to a Sexist Status Quo. *Huffington Post*. https://www.huffingtonpost.com/soraya-chemaly/female-nudity-isnt-obscen_b_5186495.html; Datta, B. (2014, 16 September). Never mind the nipples: Sex, gender and social media. *GenderIT.org*. https://www.genderit.org/feminist-talk/never-mind-nipples-sex-gender-and-social-media; and Pasricha, J. (2016, 10 November). It's 2016 and Facebook is still terrified of women's nipples. *GenderIT.org*. https://www.genderit.org/feminist-talk/its-2016-and-facebook-still-terrified-womens-nipples

[54] https://www.facebook.com/communitystandards

[55] 7amleh – The Arab Centre for Social Media Advancement. (2017). *The Index of Racism and Incitement on Israeli Social Media 2016*. 7amleh.org/2017/02/07/7amleh-center-publishes-the-index-of-racism-and-incitement-in-the-israeli-social-media-2016

[56] Nashif, N. (2017, 10 April). Facebook vs Palestine: Implicit Support for Oppression. *Al Jazeera*. www.aljazeera.com/indepth/opinion/2017/04/facebook-palestine-implicit-support-oppression-170409075238543.html

[57] The Guardian. (2017, 21 May). Facebook's Manual on Credible Threats of Violence. https://www.theguardian.com/news/gallery/2017/may/21/facebooks-manual-on-credible-threats-of-violence

[58] Sign-on letter against Facebook censorship policy: https://s3.amazonaws.com/s3.sumofus.org/images/79_FacebookCensorshipPolicySign-OnLetter.pdf

[59] See, for example, cases from Bosnia and Herzegovina and from Pakistan in Athar, R. (2015). *From impunity to justice: Improving corporate policies to end technology-related violence against women*. Association for Progressive Communications. https://www.genderit.org/sites/default/upload/flow_corporate_policies_formatted_final.pdf

Companies should notify users why content was restricted and taken down or accounts were suspended, and allow them to appeal immediately. If there is in fact a ToS violation, users should have the opportunity to revise their post to have it reposted. Users should be able to appeal in their own language and reach someone in their own time zone. Without such measures, users may continually and unintentionally repeat the violation and have their account disabled, exacerbating the violation of their freedom of expression. As part of their responsibilities under the UN Guiding Principles on Business and Human Rights,[60] companies are required to provide access to remedy in order to mitigate harm, and grievance mechanisms are critical in this regard. This also represents a missed opportunity for educating users, rather than simply taking down content. If users are violating ToS, for example through non-consensual sharing of intimate images, notifying them of the reason for the takedown is an opportunity for the platform to educate the user on its anti-violence standards and definitions.

9. Automation and content moderation

With the colossal – and growing – amount of content posted to platforms every day, it is understandable that companies look to automation to identify potentially infringing content. Most major platforms contend that they only use automation for flagging content, and that is a human reviewer that ultimately makes the decision about whether it should be taken down or not.[61] However, APC observes that even as a tool for flagging content, automation, such as algorithmic filtering, is resulting in the removal of content that does not violate ToS, and is of public interest. Some examples include:

- *War crime evidence:* Platforms are deleting content that constitutes evidence of war crimes, as it gets flagged as violent content/promoting extremism, putting into jeopardy future war crime tribunal cases.[62]
- *Algorithm-based translation:* Platforms rely on automation for translating content, which can lead to arbitrary takedowns, and even arrests.[63]

Automated content moderation can also be used to give users more control. For example, Facebook's machine learning models can recognise the content of photos, so users should be able to choose an option for "no nudity" rather than Facebook banning it wholesale. We agree with EFF that "in general platforms can and should simply use smart filters to better flag potentially unlawful content for human review and to recognize when their user flagging systems are being gamed by those seeking to get the platform to censor others."[64]

We acknowledge that it is unrealistic to expect companies to perfectly differentiate between speech that should be protected and speech that should be erased. We are not opposed to the use of artificial intelligence to identify problematic content for human review, as we are aware of how stressful the work of content moderators who have to spend hours looking at often very disturbing content can be.

[60] Ruggie, J. (2011). Op. cit.

[61] See, for example: YouTube Official Blog. (2017, 4 December). Expanding our work against abuse of our platform. https://youtube.googleblog.com/2017/12/expanding-our-work-against-abuse-of-our.html

[62] Asher-Schapiro, A. (2017, 2 November). YouTube and Facebook Are Removing Evidence of Atrocities, Jeopardizing Cases Against War Criminals. *The Intercept*. https://theintercept.com/2017/11/02/war-crimes-youtube-facebook-syria-rohingya

[63] Ong, T. (2017, 24 October). Facebook apologizes after wrong translation sees Palestinian man arrested for posting 'good morning'. *The Verge*. https://www.theverge.com/us-world/2017/10/24/16533496/facebook-apology-wrong-translation-palestinian-arrested-post-good-morning

[64] McSherry, C., York, J. C., & Cohn, C. (2018, 30 January). Op. cit.

However, when such automated processes are used it is inevitable that they will “make mistakes” and that therefore their use should be made transparent, all content removal should be subject to human review, and users should have easy recourse to challenging removals which they believe to be arbitrary or unfair.

10. Transparency

While both the number of platforms releasing transparency reports in recent years[65] and the details contained in those reports has increased, in APC's view, companies still need to do more to comply with their responsibility to ensure that their policies for restricting content are being applied in a non-discriminatory and equitable manner that provides users with access to remedy. As noted previously, companies report very little publicly about when they remove content or restrict users' accounts for violating their ToS, and targets of reporting often have no idea what rule they have violated, since companies often fail to provide adequate notice. Moreover, companies enter into agreements with states in order to operate locally, the terms of which have implications for content regulation, but are often completely unknown.

Another area where improvement is needed is around reporting of government requests received. While companies do report on the number of pieces of content restricted due to government requests, per country, they do not consistently disclose the total number of requests received. Having this additional information would make it possible to understand whether requests are on the rise, and the extent to which the company is complying with, or pushing back against, government requests, which is critical for monitoring and accountability.

In our view, companies need to:

- Refrain from entering into opaque agreements with governments regarding content regulation.
- Expeditiously notify users that their content has been restricted, on what basis they are restricting content, and avenues for appealing the decision in easily understandable language.
- Disclose information on enforcement of their ToS.
- Make public their guidelines on how ToS are implemented.

We recognise that measures will need to be taken to ensure that more transparency around ToS implementation guidelines does not further empower trolls to manipulate content removal processes to silence opposing views. There should be an opportunity for feedback from the community of users and human rights experts to ensure that implementation guidelines comply with principles of equality and non-discrimination.

11. General recommendations from APC

State responsibility: The primary responsibility of respecting, protecting and fulfilling human rights lies with the state. This includes the duty to protect against human rights abuses by third parties, including business. As such, APC believes that states should:

[65] Since 2010, when companies first started publishing transparency reports, the number of companies doing so has increased to 68, covering 90 countries. See the Access Now Transparency Reporting Index: https://www.accessnow.org/transparency-reporting-index

- Refrain from outsourcing to companies measures that constitute violations of freedom of expression, as they are doing through secretive agreements, codes of conduct, ToS reporting for alleged infringements, and similar measures.
- Adopt safeguards to ensure that when working with the private sector to take down unlawful content, they do not circumvent legal procedures.
- Be clear and transparent as to whether they are issuing a legal order requiring content removal or restriction, or submitting to a company a referral of an alleged ToS violation. Using ToS referrals as a basis for content removals or restrictions undermines due process.

Private sector accountability: All companies, regardless of their size, have a responsibility to respect human rights, by not infringing on the human rights of users and addressing adverse human rights impacts with which they are involved. This submission focuses on the large internet platforms, partly because their services are used globally by a large number of users. We also question whether these large platforms need to be treated in a different category because of the way in which they dominate the market. Aside from being very large companies, we question if there is something distinct in the way they operate, with their networked effect and the lack of alternatives, that impacts on users' exercise of freedom of expression.

While this submission focuses on the large internet platforms, APC emphasises that all internet platforms should have in place policies and practices that are appropriate for their size and resources, and also reflective of their user bases, including:

- A policy commitment to meet their responsibility to respect human rights.
- A human rights due diligence process to identify, prevent, mitigate and account for how they address their impacts on human rights.
- Processes to enable the remediation of any adverse human rights impacts they cause or to which they contribute.[66]

Human rights impact assessments: Given that companies are constantly introducing new products, updating their policies, and expanding into new jurisdictions, human rights impact assessments should be carried out on an ongoing basis, and should not be a one-time event. Human rights impact assessments should include all human rights that companies' policies may impact, beyond freedom of expression and privacy, to include also economic, social and cultural rights, the right to be free from violence, and the right to participate in public life, among others. In addition, they should consider how their policies can strengthen, rather than undermine, due process.

Due process: Every user should have the right to due process, including to be expeditiously notified of content takedowns, the reason for the takedown, and the option to appeal a company's takedown decision, in every case. The Manila Principles provide a framework for this:

- Legitimacy: The mechanism is viewed as trustworthy and is accountable to those who use it.
- Accessibility: The mechanism is easily located, used and understood. It needs to use clear understandable language (both plain and local language) reflecting user base.
- Predictability: There is a clear and open procedure with indicative time frames, clarity of process and means of monitoring implementation.

[66] Ruggie, J. (2011). Op. cit.

- Equitable: It provides sufficient information and advice to enable individuals to engage with the mechanism on a fair and informed basis.
- Transparent: Individuals are kept informed about the progress of their matter.
- Rights-respecting: The outcomes and remedies accord with international human rights principles of non-discrimination and equality.
- Source of continuous learning: The mechanism enables the platform to draw on experiences to identify improvements for the mechanism and to prevent future grievances.

Implementation of content restriction: While platforms are increasingly publishing their content restriction policies online, much more transparency is needed in terms of how they are being implemented. In particular, greater attention is needed to ensure that policies are upholding the international human rights principles of non-discrimination and equality, and are taking into account contextual factors, such as language, culture, and power dynamics. Companies should:

- Be more transparent about how they interpret and implement their content restriction policies, and facilitate a process for feedback from the community of users and human rights experts.
- Provide additional training for moderators that addresses cultural and language barriers, power dynamics, and issues such as gender bias and LGBTIQ sensitivity.
- Hire more speakers of languages that are currently under-represented among content moderators.
- Provide content reviewers with psychosocial support and resources to deal with the disturbing and harmful content they are reviewing.

Transparency: Increased transparency is needed in a number of areas in order to better safeguard freedom of expression against arbitrary content removals and to better understand how the content viewed online is being moderated:

- Companies should disclose more data in their transparency reports in the following areas: content removed through ToS enforcement, and the full number of content removal requests they receive, including requests that were rejected.
- Platforms should allow truly independent researchers access to work with, black box test and audit their systems.
- Users should be able to see what is shown in their feed and why.
- APC supports the development of a standardised public interest application programming interface (API) model, which would offer a degree of visibility to the contents and origins of deleted materials on social networks.[67] A public interest API would not force companies to reveal their proprietary algorithms or threaten the privacy of users. It would simply make it possible to understand what content is fed into the algorithm (i.e. who created an advertisement, for example) and how the algorithm distributed that content (i.e who was targeted with the advertisement).

Alternatives to taking down content: Content removals are just one way of addressing content that may be harmful to other users. Platforms are building tools that let users filter ads and other content. While this approach has the potential to further "information bubbles", it can also empower users to make informed choices about the content that they see. This may be preferable to companies making

[67] Ghonim, W., & Rashbass, J. (2017, 31 October). It's time to end the secrecy and opacity of social media. *Washington Post*. https://www.washingtonpost.com/news/democracy-post/wp/2017/10/31/its-time-to-end-the-secrecy-and-opacity-of-social-media/?utm_term=.d34a4f1b771e

these decisions for users, when at the end of the day companies’ criteria will be influenced by what content is deemed profitable. We encourage companies to explore tools that enable users to be in more control of their own online experience.

Multistakeholder process: We recommend the establishment of a multistakeholder process, building on existing multistakeholder initiatives, with input from different parts of the world, to develop global guidelines or norms to address the challenge of harmful content within a rights-respecting framework. This multistakeholder process could explore whether establishing a more traditional self-regulatory framework would have positive or negative consequences.