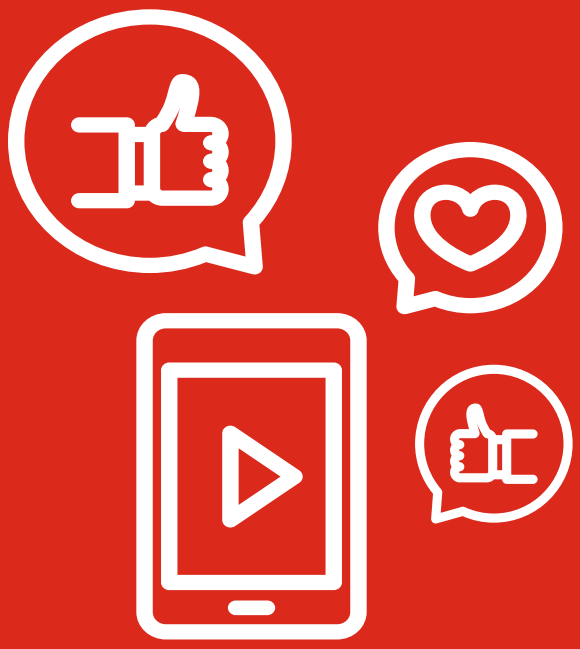
ARTICLE 19



Side-stepping rights: Regulating speech by contract

2018

Policy Brief

First published by ARTICLE 19, 2018

ARTICLE 19
Free Word Centre
60 Farringdon Road
London EC1R 3GA
UK

www.article19.org
T: +44 20 7324 2500
E: info@article19.org
Tw: @article19org
Fb: facebook.com/article19org

Printed by Marstons, England.

ARTICLE 19 works for a world where all people everywhere can freely express themselves and actively engage in public life without fear of discrimination. We do this by working on two interlocking freedoms, which set the foundation for all our work. The Freedom to Speak concerns everyone's right to express and disseminate opinions, ideas and information through any means, as well as to disagree from, and question power-holders. The Freedom to Know concerns the right to demand and receive information by power-holders for transparency good governance and sustainable development. When either of these freedoms comes under threat, by the failure of power-holders to adequately protect them, ARTICLE 19 speaks with one voice, through courts of law, through global and regional organisations, and through civil society wherever we are present.

Created in partnership with:

SWEDEN

CIVIC SPACE INITIATIVE

ARTICLE 19 · WORLD MOVEMENT FOR DEMOCRACY · CIVICUS · ICNL

"This publication is wholly financed by the Government of Sweden. The Government of Sweden does not necessarily share the opinions here within expressed. ARTICLE 19 bears the sole responsibility for the content."

Contents

Executive summary

In this policy brief, ARTICLE 19 addresses the compliance of dominant social media platforms – Facebook, Twitter, and YouTube (owned by Google) – with international freedom of expression standards; and offers some practical recommendations as to the steps companies should take in order to demonstrate their commitment to the protection of freedom of expression.

While freedom of expression has generally enjoyed high levels of protection on social media platforms, they have increasingly had to address to the human rights concerns of the various communities they seek to attract on their platforms. They are also under constant pressure from governments to remove content deemed harmful or illegal under respective national laws. Online censorship is therefore increasingly privatised. This raises serious questions for the protection of freedom of expression online.

In this policy, ARTICLE 19 puts forward that although social media companies are in principle free to restrict content on the basis of freedom of contract, they should respect human rights, including the rights to freedom of expression, privacy and due process. The policy sets out the applicable standards for the protection of freedom of expression online, particularly as it relates to social media companies, and lays down the key issues that arise in relation to the regulation of speech by contract. It further provides an analysis of selected Terms of Service of four dominant social media companies and examines the various policy options available to regulate social media platforms. Finally, ARTICLE 19 makes recommendations as to the basic human rights standards that companies should respect.

Key recommendations:

Recommendations to States

- States should adopt laws that shield social media companies from liability for third-party content and refrain from adopting laws that would make them subject to broadcasting regulatory authorities or other similar public authorities;
- States should refrain from putting undue extra-legal pressure on social media companies to remove content;
- States should provide for a right to an effective remedy for violations of freedom of expression by social media companies.

Recommendations to social media companies

- Companies should ensure that their Terms of Service are sufficiently clear, accessible and in line with international standards on freedom of expression and privacy. They should also provide more detailed examples or case studies of the way in which their community standards are applied in practice;
- Companies should be more transparent about their decision-making processes, including the tools they use to moderate content, such as algorithms and trusted flagger-schemes;
- Companies should ensure that sanctions for non-compliance with their Terms of Service are proportionate;
- Companies should put in place internal complaints mechanisms, including for the wrongful removal of content or other restrictions on their users' freedom of expression;
- Companies should collaborate with other stakeholders to develop new independent self-regulatory mechanisms;
- Companies should resist government and court orders in breach of international standards on freedom of expression or privacy;.
- Companies should publish comprehensive transparency reports, including detailed information about content removal requests received and actioned on the basis of their Terms of Service. Additional information should also be provided in relation to appeals processes, including the number of appeals received and their outcome.

Introduction

In the digital world, social media has become fundamental to how people communicate. According to recent estimates, there are currently 2.2 billion active Facebook users,[1] and 330 million Twitter users;[2] meanwhile, a billion hours of video are watched daily on YouTube.[3] Despite their extraordinarily positive effect on freedom of expression, these companies have come to wield enormous power over what information individuals have access to.

Contrary to the common perception that 'anything goes online,' sharing information and/or opinions on social media platforms is not control-free. When users join Facebook, Twitter or YouTube, they accept abiding by those companies' Terms of Service. These Terms of Service typically include community standards that lay down the types of content that the respective company deems acceptable or not. Social media users who fall foul of these standards may therefore see their content removed or their account disabled altogether.

In addition, these companies face constant pressure from governments to remove more content – from 'hate speech' and 'extremist' content to 'fake news' – or are strongly incentivised to remove content that may be in breach of the law of the country in which they operate, lest they be found liable for illegal content under 'notice-and-takedown' regimes. Online censorship is therefore increasingly privatised.

Moreover, social media companies use algorithms to prioritise users or news feeds. While this usually takes place on the basis of the perceived interests of their users, it is also the result of advertising or other marketing agreements. "In short, the vast majority of speech online is now regulated by the contractual terms of a handful of companies, mostly based in the United States (US)".

Freedom of expression used to enjoy high levels of protection within social media platforms. However, as social media platforms have grown to encompass hundreds of millions of users all over the world, the companies have had to address the human rights concerns of the various communities they seek to attract on their platforms. This raises serious questions for the protection of freedom of expression, such as:

- What free speech standards should social media companies respect?
- Given that social media companies are effectively services provided by private companies, can they be required to comply with international standards on freedom of expression?
- Does the quasi-public nature of some of these online spaces call for a different type of regulation?
- What are the minimum procedural safeguards companies should respect to ensure strong protection of freedom of expression?

This policy brief seeks to answer these and other questions in light of international standards on freedom of expression. It builds on our previous policy on *Internet Intermediaries: Dilemma of Liability* [4] and also offers some practical recommendations as to the steps companies should take in order to demonstrate their commitment to the protection of freedom of expression. The scope of our enquiry is purposefully narrow: in this brief, we focus on the basic international standards on free expression that major (or dominant) social media companies – such as Facebook, Twitter, and YouTube (owned by Google) – should respect in developing and applying their Terms of Service. The extent to which the Terms of Service of the above-mentioned companies comply with international norms is also briefly examined.[5] This brief does not address the free speech implications of Terms of Use as applied by telecommunications operators ('telcos') or other online service providers such as PayPal, Mastercard and Visa. This particular aspect of speech regulation by contract is examined in separate policy briefs.[6]

This policy brief is divided into five parts. First, it sets out the applicable standards for the protection of freedom of expression online, particularly as it relates to social media companies. Second, it lays down the key issues that arise in relation to the regulation of speech by contract. Third, it provides an analysis of selected Terms of Service of four dominant social media companies. Fourth, it examines the various policy options available to regulate social media platforms. Finally, ARTICLE 19 makes recommendations as to the basic human rights standards that companies should respect.

Applicable international standards

Guarantees of 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),[7] and given legal force through Article 19 of the **International Covenant on Civil and Political Rights** (ICCPR).[8] Similar guarantees to the right to freedom of expression are further provided in the regional treaties.[9]

The scope of the right to freedom of expression is broad. It requires States to guarantee to all people the freedom to seek, receive or impart information or ideas of any kind, regardless of frontiers, through any media of a person's choice. In 2011, the UN Human Rights Committee (HR Committee), the treaty body monitoring States' compliance with the ICCPR, clarified that the right to freedom of expression applies also to all forms of electronic and Internet-based modes of expression; and that the legal framework regulating the mass media should take into account the differences between the print and broadcast media and the Internet.[10] Similarly, the four special mandates on freedom of expression highlighted in their 2011 *Joint Declaration on Freedom of Expression and the Internet* that regulatory approaches in the telecommunications and broadcasting sectors could not simply be transferred to the Internet.[11] In particular, they recommended the adoption of tailored approaches to address illegal content online, while pointing out that specific restrictions for material disseminated over the Internet were unnecessary.[12] They also encouraged the promotion of self-regulation as an effective tool in redressing harmful speech.[13]

Limitations on the right to freedom of expression

Under international human rights standards, States may, exceptionally, limit the right to freedom of expression provided that such limitations conform to the strict requirements of the three-part test. This requires that limitations must be:

- **Provided for by law.** 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; and
- **Necessary and proportionate in a democratic society**, requiring that if a less intrusive measure is capable of achieving the same purpose as a more restrictive one, the less restrictive measure must be applied.[14]

Further, Article 20(2) of the 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.[15]

Social media companies and freedom of expression

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

Intermediary liability

The four special mandates on freedom of expression have recognised for some time that immunity from liability was the most effective way of protecting freedom of expression online. For example, in their 2011 Joint Declaration, they recommended that intermediaries should not be liable for content produced by others when providing technical services, and that liability should only be incurred if the intermediary has specifically intervened in the content, which is published online.[16]

In 2011 the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (Special Rapporteur on FOE) stated that censorship should never be delegated to a private entity, and that States should not use or force intermediaries to undertake censorship on its behalf.[17] 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.[18]

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 Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework* (the Guiding Principles) provide a starting point for articulating the role of the private sector in protecting human rights on the Internet.[19] The Guiding Principles 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:[20]

    - Make a public statement of their commitment to respect human rights, endorsed by senior or executive-level management;

    - Conduct due diligence and human rights impact assessments in order to identify, prevent and mitigate against any potential negative human rights impacts of their operations;

- Incorporate human rights safeguards by design in order to mitigate adverse impacts, and build leverage and act collectively in order to strengthen their power vis-a-vis government authorities;
- Track and communicate performance, risks and government demands; and
- Make remedies available where adverse human rights impacts are created.

- In his May 2011 report to the Human Rights Council, the Special Rapporteur on FOE highlighted that – while States are the duty-bearers for human rights – Internet intermediaries also have a responsibility to respect human rights and referenced the Guiding Principles in this regard.[21] 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.[22] 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.[23] 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 on the potential pitfalls of their misuse.[24]

- In his June 2016 Report to the Human Rights Council,[25] 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,"[26] 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."[27]

- 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 Guiding Principles[28] 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.[29] He also encouraged companies to publish transparency reports about government requests for user data or content removal;[30] challenge requests for content removal or requests for user data that may violate the law or internationally recognised human rights;[31] notify individuals affected by any measure restricting their freedom of expression and provide them with non-judicial remedies;[32] and take proactive protective measures to develop good business practices consistent with respect for human rights.[33]

- In the 2016 report on Standards for a Free, Open and Inclusive Internet,[34] the OAS Special Rapporteur on FOE recommended that, *inter alia,* companies make a formal and high-level commitment to respect human rights, and back this commitment up with concrete internal measures and systems; seek to ensure that any restriction based on companies' Terms of Service do not unlawfully or disproportionately restrict freedom of expression; and put in place effective systems of monitoring, impact assessments, and accessible, effective complaints mechanisms.[35] He also highlighted the need for companies' policies, operating procedures and practices to be transparent.[36]

- At European level, in an *Issue Paper on the Rule of law on the Internet* and in the wider digital world, the Council of Europe Commissioner for Human Rights recommended that States should stop relying on private companies that control the Internet to impose restrictions that violate States' human rights obligations.[37] He recommended that further guidance should be developed on the responsibilities of business enterprises in relation to their activities on (or affecting) the Internet, in particular to cover situations in which companies may be faced with demands from governments that may be in violation of international human rights law.[38]

- Similarly the Committee of Ministers of the Council of Europe, in its Recommendation on the protection of human rights with regard to social networking services, recommended that social media companies should respect human rights and the rule of law, including procedural safeguards.[39] Moreover, in its March 2018 Recommendation on **the roles and responsibilities of internet intermediaries**, the Committee of Ministers adopted detailed recommendations on the responsibilities of Internet intermediaries to protect the rights to freedom of expression and privacy and to respect the rule of law.[40] 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.[41] In particular, they make clear that companies' content restriction practices must comply with the tests of necessity and proportionality under human rights law;[42] and that intermediaries should provide users with complaints mechanisms to review decisions to restrict content made on the basis of their content restriction policies.[43]

- Similarly, the Ranking Digital Rights project has undertaken a ranking of the major Internet companies by reference to their compliance with digital rights indicators. These include the following freedom of expression benchmarks: (i) availability of Terms of Service; (ii) terms of service, notice and record of changes; (iii) reasons for content restriction; (iv) reasons for account or service restriction; (v) notify users of restriction; (vi) process for responding to third-party requests; (vii) data about government requests; (viii) data about private requests; (ix) data about Terms of Service enforcement; (x) network management (telecommunication companies); (xi) identity policy (internet companies).[44]

- Finally, the Dynamic Coalition on Platform Responsibility is currently seeking to develop standard Terms and Conditions in line with international human rights standards.[45]

Content-specific principles

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

- The 2016 *Joint Declaration on Freedom of Expression and Countering Violent Extremism* recommends that States should not subject Internet intermediaries to mandatory orders to remove or otherwise restrict content, except where the content is lawfully restricted in accordance with international standards.[46] Moreover, they recommended that any initiatives undertaken by private companies in relation to countering violent extremism should be robustly transparent, so that individuals can reasonably foresee whether content they generate or transmit is likely to be edited, removed or otherwise affected, and whether their user data is likely to be collected, retained or passed to law enforcement authorities.[47]

- The 2017 *Joint Declaration on 'Fake news', Disinformation and Propaganda* recommended, *inter alia,* that intermediaries adopt clear, pre-determined policies governing actions that restrict third party content (such as deletion or moderation) which goes beyond legal requirements.[48] 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.[49] Intermediaries should also take effective measures to ensure that their users can both easily access and understand their policies and practices,

including Terms of Service, including detailed information about how they are enforced, and, where relevant, by making clear, concise and easy to understand summaries of, or explanatory guides to, those policies and practices, available.[50] 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.[51]

- 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.[52] 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, as well as those who experience multiple and intersecting discrimination.[53] 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.[54]

The inability to communicate privately substantially affects individuals' freedom of expression rights. In his report of May 2011, the Special Rapporteur on FOE expressed his concerns over the fact that States and private actors use the Internet to monitor and collect information about individuals' communications and activities on the Internet, and that these practices can constitute a violation of Internet users' right to privacy, and ultimately impede the free flow of information and ideas online.[55]

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.[56]

Further, in his May 2015 report on encryption and anonymity in the digital age, the Special Rapporteur on FOE recommended that States refrain from making the identification of users a pre-condition for access to digital communications and online services, and from requiring SIM card registration for mobile users.[57] He also recommended that corporate actors reconsider their own policies that restrict encryption and anonymity (including through the use of pseudonyms).[58]

Regulating speech by contract: the problems

As already noted, social media companies have come to hold exceptional influence over individuals' exercise of their right to freedom of expression online. ARTICLE 19 finds that the privatisation of speech regulation (i.e. the regulation of speech by contract) raises serious concerns for the protection of freedom of expression, in particular those set out below.

Lack of transparency and accountability

Although social media companies have made some progress in transparency reporting over recent years, there remains a significant **lack of transparency and accountability** in the manner in which they implement their Terms of Service.[59] For instance, neither Facebook,[60] nor Google[61] currently publish information about content removals on the basis of their Terms of Service in their Transparency Reports. As a result, it is difficult to know if the Terms of Service and Community Standards are applied differently from country to country.

Equally, it is generally unclear whether they remove content on their own initiative, for instance because a filter or other algorithm has flagged a particular key word (for example, swear words) or video content (for example, copyright),[62] or because a takedown request has been filed by a trusted third party.[63] Although companies have recently become more open about their use of "hashes," algorithms and filters to takedown 'terrorist' video content even before it is posted, the criteria they use for such removals and how the algorithms or filters operate in practice remain uncertain.[64] Finally and in any event, given that these companies' Terms of Service are often coined in broad terms, it is generally impossible to know whether they are applied reasonably, arbitrarily or discriminatorily, bar press coverage[65] or public campaigns conducted by affected individuals or groups.[66]

Lack of procedural safeguards

There are insufficient **procedural safeguards** in the removal of content on social media. While Facebook, Twitter and YouTube generally allow their users to report content which they believe to be illegal, in breach of their community guidelines, or simply harmful, there are no obvious appeals mechanisms for users to challenge a content removal decision made by these companies.[67] The only exception to this appears to be YouTube's copyright dispute mechanism.[68]

Moreover, it is generally unclear whether or not companies notify users that their content has been removed, flagged, or their account penalised, and the reasons

for such actions. For instance, Twitter explains that it will attempt to send notice about any "legal" request that it receives, to the email address associated with the respective account.[69] However, its explanation in relation to enforcement of its Terms of Service strongly suggests that individuals whose content is taken down on that basis are not given any reasons for the takedown decision, let alone the opportunity to challenge a takedown request *before* any sanctioning measures are applied.[70]

More generally, it is unclear whether such notification takes place on a systematic basis or whether affected users have the opportunity to challenge decisions made on the basis of their alleged violation of the platform's Terms of Service, *after* sanctioning measures have been applied to their account. The Terms of Service also do not contain any clear information about redress mechanisms for wrongful removal of content.

Lack of remedy for the wrongful removal of content

Another critical shortfall of social media platforms is a **lack of remedy for the wrongful removal of content** on the basis of companies' Terms of Service. The same is true in relation to legal remedies. Whereas individuals aggrieved by information posted on social media or the Internet at large can usually rely on legal remedies including, for example, harassment, defamation, the misuse of private information, or, more recently, the so-called 'right to be forgotten', there are no such remedies available to individuals whose content is wrongfully removed as a result of these complaints.

Unfair contract terms

Terms of Service are generally formulated in such a way as to create **imbalances of power** and **unfair contract terms** between the companies and individuals.[71] Moreover, since Terms of Service typically stipulate a jurisdiction for dealing with any contractual claim or dispute arising from the use of the service,[72] such terms are likely to create significant **barriers to access to justice** for those users based outside of the respective jurisdiction.[73] Some of these imbalances are slowly being redressed: Google, Twitter and Facebook recently yielded to the demands of EU consumer watchdogs and, among other things, accepted making their jurisdiction clauses compliant with EU law.[74] Notwithstanding these improvements, users are likely to find it difficult to resist the broad immunity granted to intermediaries under section 230 of the Communications Decency Act (CDA) for any action taken voluntarily and in good faith to restrict access to material that they consider objectionable.[75] Similarly, US jurisprudence suggests that Facebook or YouTube's Terms of Service are unlikely to be found unconscionable, nor termination of user accounts in breach of good faith and fair dealing implied terms, save in exceptional circumstances.[76] In Europe, judges have eschewed ruling on the wrongful removal of content on the basis of companies' Terms of Service.[77]

Lower free speech standards

Terms of Service usually include **lower standards for restrictions on freedom of expression than those permitted under international human rights law**.[78] While low free speech standards generally enable companies to grow their user-bases by creating safer online environments, they also turn these quasi-public spaces into much more sanitised environments, in which freedom of expression is not limited by the principles of necessity and proportionality but rather by propriety.

In practice, low free speech standards are often the result of companies adapting their community standards to domestic legal requirements that fall below international standards on freedom of expression. For instance, in 2015, both Twitter and Facebook updated their content policies to prohibit the "promotion" of violence or terrorism.[79] In particular, Facebook Community Standards now provide that "supporting" or "praising" leaders of terrorist organisations, or "condoning" their violent activities, is not permitted.[80] This, however, falls below international standards on freedom of expression, which only allow the prohibition of incitement to terrorism or violence rather than their glorification or promotion.

Low free speech standards are not limited to finding the lowest common denominator between the myriad of laws that social media companies may be required to comply with. They may also be driven by the demands of the advertising industry, which do not want their image to be tarnished by being associated with problematic content. For instance, in 2013, Facebook was widely criticised for allowing images glorifying rape to be posted on their platform and a number of advertisers (including Dove, Nationwide and Nissan) pulled their advertising over the policy.[81] Facebook eventually removed the images from its platform and its policy now says that Facebook removes content "that threatens or promotes sexual violence."[82] More recently, Google apologised to advertisers for 'extremist' content appearing on YouTube.[83] This was followed by announcements that the company was stepping up its efforts and using artificial intelligence to stamp out 'extremism' from its services, notwithstanding the ambiguity of this term.[84]

Circumventing the rule of law

Finally, public authorities, and, in particular, law enforcement agencies, regularly seek the cooperation of social media platforms with a view to combat criminal activity (e.g. child pornography) or other social harms (e.g. 'online extremism' in a way which **circumvents the rule of law**.[85] In particular, because these authorities do not always have the power to order the removal of the content at issue, they contact social media platforms informally and request the removal of content on the basis of the companies' Terms of Service. While in principle the companies are not required to comply with such requests in the absence of a judicial determination of the legality of the content at issue, it puts them in a difficult position in circumstances where the content may be at the fringes of illegality. The net result is that social media companies often become the long arm of the law without users being afforded the

opportunity to challenge the legality of the restriction at issue before the courts. This problem is even more acute when these forms of cooperation or public-private partnerships are put on a legal footing or promoted as part of trade agreements.[86] In reality, this only serves to institutionalise private law enforcement by default.

Analysis of Terms of Service of dominant social media companies

To demonstrate the outlined problems with content removal, ARTICLE 19 analyses selected aspects of the Terms of Service of the dominant global social media companies – Google, YouTube, Twitter and Facebook. Since a detailed analysis of their entire Terms of Service is beyond the scope of this policy, we examine content restrictions in the areas that are most often found to be problematic, in particular 'hate speech',[87] 'terrorist' and related content, so-called 'fake news,' as well as privacy and morality-based restrictions on content. We further examine the procedural issues related to the removal of content on the basis of the respective Terms of Service.

Content restrictions

Hate speech

'Hate speech'[88] is a major of issues for social media, particularly as they strive to provide 'safe' environments for their users and are pressured by governments and the public to address 'hate' online.[89] Accordingly, they have all adopted policies to deal with 'hate speech' with varying degrees of precision. While it is legitimate for companies to seek to address 'hate speech' concerns, ARTICLE 19 finds that their policies overall fall below international standards on freedom of expression.

Facebook

Facebook's Community Standards[90] stipulate that Facebook does not allow 'hate speech' on its platform.[91] It defines 'hate speech' as "a direct attack on people based on what we call protected characteristics – race, ethnicity, national origin, religious affiliation, sexual orientation, sex, gender, gender identity and serious disability or disease."[92] Some protections are provided for "immigration status."[93]

According to the Community Standards related to 'hate speech,' Facebook interprets 'hate speech' as:

- violent speech as including support for death/disease/harm;
- dehumanising speech as including reference or comparison with filth, bacteria, disease or faeces, reference or comparison to sub-humanity, reference or comparison with animals that are culturally perceived as physically or intellectually inferior;

- statements of inferiority as including a statement or term implying a person's or a group's physical, mental or moral deficiency, statements of contempt;[94] expressions of disgust or cursing at people who share a protected characteristic.

- Content that describes or negatively targets people with slurs is also not allowed.[95]

ARTICLE 19 welcomes that the 2018 version of the Community Standards is far more detailed than previous iterations. In particular, it lays down more criteria for content restrictions than before. However, these criteria are still far broader than those permitted under international law. For instance, "attack" is broadly defined as encompassing "violent speech," "dehumanising statements" or "statements of inferiority" without making any reference to either the intent of the speaker to incite others to take action, or the likelihood of a specific type of harm occurring as a result of the speech at issue. The examples given suggest that many different types of legitimate speech are likely to be removed.

Although Facebook understandably seeks to create a 'safe' environment for its users, it effectively sets a very low bar for free expression, where views deemed offensive are likely to be removed. Moreover, the lack of sufficiently detailed examples or case studies of how the standards should be applied in practice, means that it is highly likely that Facebook's application of its policies will continue to be arbitrary and biased.

The Community Standards further state that Facebook does not allow any organisations or individuals that are engaged in organised hate on its platform.[96] "Hate organisations" are defined as "any association of three or more people that is organised under a name, sign or symbol and which has an ideology, statements or physical actions that attack individuals based on characteristics, including race, religious affiliation, nationality, ethnicity, gender, sex, sexual orientation, serious disease or disability."[97] ARTICLE 19 finds that this definition is incredibly broad and could arguably include some political parties, though the threshold that would need to be reached for such organisations to be banned remains unclear. The term "attack" remains unclear and there is no mention of intent to harm particular groups. Coupled with a broad ban on "content that praises any of the above organisations or individuals or any acts committed by the above organisations or individuals" and on "co-ordination of support for any of the above organisations or individuals or any acts committed by the above organisations or individuals," it seems inevitable that content considered legitimate under international law will get caught.

The prohibitions contained in the section on 'Dangerous individuals and organisations' provide for no exceptions. However, the 'Hate speech section' explains that sharing content containing 'hate speech' for the purposes of raising awareness and educate others is allowed, although users are expected to make their intent in this regard clear, or the content might be removed. Humour and social commentary are also allowed. Whilst these exceptions are welcome, they are unduly limited. In particular,

the requirement that users make explicit their intent to educate others sets a high threshold for the exceptions to apply. It seems unrealistic and unlikely that users will state their intent explicitly whenever they comment or joke about an issue, for example; and, individuals also will not always share the same sense of humour. Moreover, there is no guidance, such as in the form of specific examples, on how Facebook applies these standards in practice.[98]

Twitter

Twitter does not use the term 'hate speech' but refers to "hateful conduct" as one of several types of prohibited abusive behaviour.

The Twitter Rules provide that: "[users] may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories."[99] The Twitter 'Hateful Conduct policy' further defines "hateful conduct" as including "violent threats; wishes for the physical harm, death, or disease of individuals or groups; references to mass murder, violent events, or specific means of violence in which/with which such groups have been the primary targets or victims; behaviour that incites fear about a protected group; repeated and/or or non-consensual slurs, epithets, racist and sexist tropes, or other content that degrades someone."[100]

ARTICLE 19 finds that the Twitter rules are extremely broad and go far beyond international standards. Although Twitter highlights the importance of context in making decisions about the enforcement of its policies, it does not explain what a consideration of context might include. For example, there is no reference to parody or humour as exemptions.[101]

Additionally, in 2018 Twitter clarified its 'Abusive Profile Information' policy; it now prohibits its users from using their "username, display name, or profile bio to engage in abusive behaviour, such as targeted harassment or expressing hate towards a person, group, or protected category."[102] It states that Twitter will review and take enforcement action against accounts that engage in "violent threats; abusive slurs, epithets, racist, or sexist tropes; abusive content that reduces someone to less than human; content that incites fear," through their profile information. As with the 'hateful conduct' policy, this goes beyond international standards, and does not explicitly provide for any exceptions in the case of social commentary, humour, or parody.

Overall, while Twitter emphasises that context is important in the enforcement of its policies, it regrettably does not provide any examples of how the policies are implemented in practice.

Youtube

YouTube Community Guidelines provide that although YouTube tries to defend a right to express unpopular points of view, it does not permit 'hate speech' on its platform.[103]

'Hate speech' is defined by YouTube 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."[104] YouTube also recognises 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."[105] It further provides that "not everything that's mean or insulting is hate speech."[106] ARTICLE 19 notes that while these caveats are positive in clarifying that content should not be restricted solely on the basis that it is offensive, the meaning of terms such as "hateful" and the circumstances in which an insult may amount to incitement to violence, hostility or discrimination, or amount to discriminatory threats or harassment, remain unclear.

Google

Google's User Content and Conduct Policy provides that it does not "support content that promotes or condones violence against individuals or groups based on race or ethnic origin, religion, disability, gender, age, nationality, veteran status or sexual orientation/gender identity, or whose primary purpose is inciting hatred on the basis of these core characteristics. This can be a delicate balancing act, but if the primary purpose is to attack a protected group, the content crosses the line."[107] ARTICLE 19 notes that this type of language is overbroad and may result in the removal of lawful expression under international law. In particular, the act of 'condoning' violence falls below the threshold of 'incitement' to violence provided under international law.

'Terrorist' and 'extremist' content

In the aftermath of a series of terrorist attacks in recent years, social media companies, have been under intense pressure from governments "to do more" to address 'terrorist' and/or 'extremist' content. ARTICLE 19 notes that the applicable policies of social media companies in this area typically fall below international standards on freedom of expression, as they use overbroad language and fail to provide any real insight into the way in which these rules should be applied.

Facebook

Facebook's Community Standards provide that Facebook does not "allow any organisations or individuals who engage in terrorist organisations, organised hate, mass or serial murders, human trafficking, organised violence or criminal activity."[108] Facebook also provides that content that "expresses support or praise for groups, leaders or individuals involved in these activities" will be removed from the platform.[109]

ARTICLE 19 notes that excluding individuals or organisations engaged in terrorist

activity from Facebook is understandable and not necessarily an unreasonable restriction on freedom of expression in and of itself. However, we also note that the lack of an agreed definition of terrorism at international level presents a key difficulty in this regard.[110] Although Facebook's definition of terrorism[111] contains a number of positive elements, such as an explicit reference to "premeditated acts of violence," with the purpose of "intimidating a civilian population, government, or international organisation", and narrows the motivations of these actors to "political, religious or ideological aims", it could be more narrowly defined. For instance, the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism (Special Rapporteur on Counter-Terrorism) suggested that any definition of terrorism should include a reference to the "intentional taking of hostages," "actions intended to cause death or serious bodily injury to one or more members of the general population or segments of it" or "actions involving lethal or serious violence to one or more members of the general population or segments of it."[112]

Facebook's definition leaves several questions unanswered, for instance whether an organisation launching a cyber-attack on critical infrastructure would fall within its definition of terrorist activity. This is compounded by the lack of explicit examples given of who or what falls in the definition of "terrorist organisation". For instance, Facebook recently designated the Arakan Rohingya Salvation Army (ARSA) as a "dangerous organisation," leading to the deletion of posts made by Rohingya civilians fleeing alleged ethnic cleansing operations by the military, on the basis of their support for ARSA.[113] Facebook explained that this was an error. However, better guidance might have helped prevent this problem.

More generally, it is unclear how Facebook deals with individuals designated as 'terrorists' by certain governments, but who may be regarded as freedom fighters, or a social movement (such as indigenous groups), with legitimate claims, by others. Although in conferences, Facebook has stated that it complies with the US State Department list of designated terrorist groups, this is not made explicit in the Community Standards. Moreover, their compliance with this list may be problematic, as it includes groups that are not designated as terrorists by the UN, such as the Kurdistan Workers' Party (PKK).

Furthermore, Facebook's policy of banning content that "expresses support" or "praise" for those groups, leaders or individuals involved in [terrorist] activities" is vague and overbroad, and inconsistent with international standards on freedom of expression.[114] In particular, and as the international mandates on freedom of expression and counter-terrorism have highlighted, any prohibition on incitement to terrorism should avoid references to vague terms such as the "promotion" or "glorification" of terrorism. For online speech to amount to incitement to terrorism, there should be an objective risk that the act incited will be committed, as well as intent that the speech at issue would incite the commission of a terrorist act.[115]

Finally, it is worth noting that Facebook can ban terrorist content under its rules on

"graphic violence"; which provides that Facebook may remove content that "glorifies violence or celebrates the suffering or humiliation of others because it may create an environment that discourages participation."[116] At the same time, Facebook notes that "people value the ability to discuss important issues such as human rights abuse or acts of terrorism." As such, it allows graphic content "to help people raise awareness about issues" but adds a "warning label to especially graphic or violent content," among other things to prevent under-18s from having access to that content.[117] Although the Community Guidelines are relatively specific on this issue, it is difficult to know how they are applied in practice in the absence of case studies. For instance, the Community Guidelines suggest that Facebook would allow beheading videos to be shared when they have journalistic or educational purposes, despite the fact that they may have been disseminated by terrorist groups in the first place. This would be in keeping with international standards on freedom of expression, but there is no data available (such as in their transparency report) to confirm that this is borne out in practice.

Twitter

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

- First, the Twitter Rules prohibit the use of Twitter "for any unlawful purposes or in furtherance of illegal activities."[118] Users "may not make specific threats of violence or wish for the serious physical harm, death, or disease of an individual or group of people" which includes, *inter alia*, "threatening or promoting terrorism."[119] Users are also prohibited from affiliating themselves with organisations that – whether by their own statements or activities on or off the platform – "use or promote violence against civilians to further their causes."[120]

- Second, the section on "violent threats and glorification of violence" in the Twitter Rules provides that Twitter "will not tolerate behaviour that encourages or incites violence against a specific person or group of people;"[121] and that it will "also take action against content that glorifies acts of violence in a manner that may inspire others to replicate those violent acts and cause real offline danger, or where people were targeted because of their potential membership in a protected category."[122] The "glorification" of terrorist attacks, rape, sexual assault and mass murders are given as examples of violations of Twitter's policy.

- Finally, Twitter defines "violent extremist groups" as groups that: (i) identify through their stated purpose, publications, or actions, as an extremist group; (ii) have engaged in, or currently engage in, violence (and/or the promotion of violence) as a means to further their cause; (iii) target civilians in their acts (and/or promotion) of violence.[123] Users will be deemed affiliates of terrorist groups if they: (i) state or suggest that an account represents or is part of a

violent extremist group; (ii) provide or distribute services (e.g., financial, media/ propaganda) in furtherance of progressing a violent extremist group's stated goals; (iii) engage in or promoting acts for the violent extremist group; and (iv) recruit for the violent extremist group.[124]

ARTICLE 19 notes that although Twitter makes an attempt at explaining its understanding of several relevant terms, it nevertheless uses vague and overbroad language such as "glorification" or "violent extremism" and generally fails to give concrete examples of how these standards are applied in practice. In particular, it is unclear how the threshold of likelihood of violence occurring is taken to consideration. Similarly, the policy on "violent extremist groups" is so broadly drafted as to catch individuals and groups that may be considered as freedom fighters, or social movements; Nelson Mandela and the African National Congress would arguably have been barred from the platform during the Apartheid regime. It is particularly problematic that the policy focuses on the designation of a speaker as a terrorist, rather than that the context and likely consequences of the speech at issue; this is contrary to international standards on freedom of expression, which do not require automatic bans on statements by terrorist and other violent groups, as long as those statements do not incite to violence.[125] Taken together, it is therefore highly likely that content, which is legitimate under international human rights law, is removed from Twitter.

Youtube

YouTube, in the section on 'Violent or Graphic Content' of their Community Guidelines, "strictly prohibits content intended to recruit for terrorist organizations, incite violence, celebrate terrorist attacks or otherwise promote acts of terrorism."[126] Additionally, it does not permit terrorist organisations to use YouTube. However, "content intended to document events connected to terrorist acts or news reporting on terrorist activities may be allowed on the site with sufficient context and intent. However, graphic or controversial footage may be subject to age-restrictions or a warning screen."[127]

ARTICLE 19 finds that YouTube's policies on terrorist content suffer from many of the shortcomings identified with regards to Twitter and Facebook. In particular, the wording is overly broad and is likely to enable the removal of legitimate content under international law. However, the exceptions to the restrictions are clearer and suggest a more nuanced approach to content removal than on Facebook and Twitter.

Google

Google enjoins its users not to use its services "to engage in illegal activities or to promote activities that are dangerous and illegal, such as terrorism."[128] It also warns that it "may also remove content that violates applicable local laws."[129] Google's rules therefore remain vague when it comes to producing 'terrorist' search results. While most large Internet companies comply with local laws, it is important to remember

that the laws of a number of countries are themselves overly broad when it comes to banning 'terrorist' or 'extremist' content, in breach of international standards on freedom of expression.

Privacy and morality-based restrictions on content

Social media companies also prohibit various types of content that would constitute an interference with the right to privacy or a restriction on freedom of expression on the grounds of public morals. This content generally falls within three categories: threats of violence, nudity or pornography and the posting of private information. ARTICLE 19 notes that 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 they constitute an interference with the right to privacy but may be otherwise justified in the public interest.

With regards to **threats of violence, harassment, online abuse and offensive content**, ARTICLE 19 makes the following observations:

- Content policies related to threats of violence tend to be drafted in relatively broad language (e.g. Twitter)[130] and generally do not include a requirement of intent, which is inconsistent with international standards on freedom of expression. At the same time, it is worth bearing in mind that threats of violence are usually criminalised so it is possible that content policies in this area are drafted broadly to encapsulate various legal requirements and make it easier for companies to deal with them under their Terms of Service rather than through the criminal law.

- "Threats of violence" is also likely to cover certain "harassment" cases, particularly for those companies that do not have specific harassment policies in place, such as YouTube.[131] Twitter's position on threats and harassment can be confusing as threats of violence generally fall under the heading of "violence and physical harm" whereas "harassment" is dealt with under "abuse", which Twitter does not define. The Twitter Rules also no longer make explicit reference to offensive content and do not explain how offensive content might be distinguished from harassment or other "hateful conduct".

- Facebook has the most detailed policy on these issues.[132] It provides an explanation of the factors it takes into account when assessing threats (including, for example, the physical location of the parties involved) and has a relatively comprehensive policy on harassment. At the same time, aspects of its policy fall below typical legal standards of harassment, since it includes breaches of its bullying policy, which is very broadly defined. As such, Facebook is likely to take measures against users in circumstances where their conduct would not be unlawful under domestic law (at least in

the absence of intent to intimidate). It would be preferable if the Community Guidelines explained in greater detail the relationship between "threats," "harassment" and "online abuse" or "bullying" and distinguished these from "offensive content", which should not be limited. They should also provide additional information about the way in which Facebook will assess "threats", "harassment" and "online abuse" or "bullying" cases, in particular to ensure protection for minority groups, and groups subject to discrimination. Vigilance is needed to ensure such restrictions are not used to wrongly censor counter-speech against harassment and 'hate speech.'

- Finally, it is important for companies to highlight that offensive content should not be taken down as a matter of principle but, on the contrary, should be allowed unless it violates other rules. Whilst it is open to companies to take down purely abusive or even offensive content, this should not be at the detriment of public debate, particularly on matters of public interest.

Pornography, nudity, non-consensual images and graphic content are generally banned by most social media platforms. ARTICLE 19 makes the following particular observations on companies' approach to these types of content, however:

- Facebook prohibits the "adult display of nudity and sexual activity."[133] However, it is more specific in relation to the types of content that are not allowed on its platform. Whilst Facebook implicitly acknowledges that its policy on nudity may be considered restrictive, it makes allowance for the posting of content depicting nudity for educational, humours or artistic purposes.[134] ARTICLE 19 generally welcomes Facebook's more detailed policies on nudity as well as the rationale it provides for its approach, including, for example, that it wishes to prevent the sharing of non-consensual intimate images. Nonetheless, we note that it appears that these restrictions are chiefly motivated by a desire to "protect" users generally from seeing certain forms of sexualised content. A key concern in this area is the lack of clarity as to how these "morality"-based terms may be enforced discriminatorily against sexual expression by women and/or lesbian, gay, bisexual and transgender (LGBT) persons. Decisions to remove such content often appear to be inconsistent with the treatment of analogous expression by cis-gendered men, or heterosexual people.

- Google bans "nude" or "sexually explicit" images in broad terms. Further, YouTube prohibits sexually explicit content such as pornography on its platform; it further makes reference to videos containing nudity or sexual content but whose primary purpose is documentary or scientific.[135]

- The Twitter Rules prohibit "unwanted sexual advances."[136] Whilst Twitter allows some adult content subject to strict restrictions under its media policy,[137] it also bans "intimate photos or videos that were taken or distributed without the subject's consent". Overall, the rules appear to be broadly similar, covering

content that may be both lawful (pornography) and unlawful in some countries (such as the non-consensual sharing of intimate images, under so-called "revenge pornography" laws). Although it is legitimate for companies to decide not to tolerate lawful content such as pornography on their platforms, the rules tend to be coined in broad terms with relatively narrow exceptions for freedom of expression. Moreover, the rules are sufficiently vague as to make it difficult to predict how Internet companies would apply them in difficult cases.

Other personal or private information, such as ID or social security information, credit card numbers, and personal medical records may also be removed by Google[138] and Twitter,[139] no doubt reflecting the type of issues faced by these Internet companies. In addition, Google has a process in place to deal with "right to be forgotten" de-listing request.[140] These rules generally appear to be in keeping with data protection and common-sense rules for the protection of the right to privacy.

In short, although community standards in this area are generally drafted in relatively plain language, they tend to ban broader categories of content than those permitted under international standards on freedom of expression. Whilst ARTICLE 19 recognises that companies might legitimately restrict access to some lawful content, such as pornography, because of the type of service they want to provide, the main problem is that their Terms of Service are drafted in overbroad terms, giving companies excessive flexibility to interpret their rules. This results in inconsistent and seemingly biased outcomes, disproportionately impacting on expression by minority groups, or groups subject to discrimination. In the absence of more concrete examples being provided within the Terms of Service or Community Guidelines of how the guidelines are applied, it is difficult to know what content actually gets removed from these platforms.

'Fake news'

Since the 2016 Presidential election campaign in the United States of America (US), governments and the public have been increasingly concerned about the dissemination of so-called 'fake news.' In response, some of the dominant social media companies have adopted initiatives ostensibly designed to combat its spread, or influence.

Facebook

Facebook's Community Guidelines state that Facebook does not allow the use of inaccurate or misleading information in order to collect likes, followers or shares as part of its anti-spam policy.[141] Further, it removes profiles that impersonate other people.[142] Facebook also started working with fact-checking organisations in order to put in place a 'fake news' labelling system,[143] under which users are able to alert Facebook to potentially false stories. In its original form, this initiative worked as follows: if enough people reported that story as 'fake,' it was then sent to trusted third-party fact-checkers. If the story was deemed by these trusted third-parties to be

unreliable, it became publicly flagged as "disputed by third-party fact checkers" and a warning appeared when users shared it.[144] Following criticism that this system was not effective,[145] Facebook decided to replace public flags with "related articles" to provide more context to stories reported as inaccurate.[146]

In 2018, Facebook announced that it would seek to tackle 'fake news' and so-called 'clickbait' by no longer prioritising pages or public content in its News Feed. Instead, it would prioritise content published by family and friends[147] or content rated as trustworthy by the Facebook community.[148] It said that more emphasis would be placed on news that people found informative and relevant to their local community.[149] In the latest version of its Community Guidelines, Facebook states that it "doesn't remove false news," but that it "significantly reduces its distribution by showing it lower in News Feed."[150] It does so, among other things, "using various signals, including feedback from our community, to inform a machine learning model that predicts which stories may be false."[151] At the same time, Facebook has stated that it has been hunting down "fake accounts" and worked with governments and civil society to defend its platform from "malicious interference," particularly during elections.[152] Facebook also claim to be strengthening enforcement of its ad policies[153] and continuing to create new tools to help its users better understand the context of the articles in its news feed.[154]

Youtube

YouTube does not ban 'fake news' *per se*. However, its Community Guidelines make clear that spam, deceptive practices and scams have no place on the platform.[155] Misleading metadata about a video, such as misleading tags, titles or thumbnails to boost viewings, can lead to the removal of content. In addition, YouTube has pledged to offer trainings to teenagers to help them identify 'fake' videos.[156] In October 2017 it was reported that YouTube was looking to change its search algorithm to produce more authoritative sources in response to searches. It remains unclear how YouTube would determine which sources are more "authoritative" and whether or not this change has been implemented.[157]

Twitter does not explicitly ban 'fake news' on its platform but a number of its policies, on impersonation,[158] spam,[159] and bots[160] may be applicable. Whilst Twitter has stated that it does not wish to be an arbiter of truth, it has recently stepped up its crackdown on some Russian "fake" accounts that allegedly interfered in the US election.[161] These efforts rely on the company's internal "systems" to detect "suspicious" activity on the platform, including suspicious accounts, tweets, logins and engagement. The company does not disclose how these "systems" are used in order to prevent "bad actors" from gaming the system.

Google

Google has not made any changes to its policies in response to 'fake news.'[162] It prohibits ads or destinations that intend to deceive users by excluding relevant information or giving misleading information about products, services, or businesses. In early 2017, Google announced that it had banned nearly 200 publishers from its ad-network, AdSense.[163] It effectively seeks to ban sites that spread misinformation from its ad-network in order to stop them from benefiting from such information. Also, in 2018, Google announced it would support the media industry by fighting misinformation and bolstering journalism, under the Google News Initiative, as a part of efforts "to work with the news industry to help journalism thrive in the digital age."[164]

ARTICLE 19 notes that the absence of an outright ban on misinformation in companies' Terms of Service is to be welcomed. Equally, companies' voluntary initiatives aimed at identifying 'fake news' in cooperation with fact-checkers are a positive step, insofar as they do not generally involve the removal of information. Nonetheless, voluntary initiatives aimed at identifying 'fake news' are a work in progress. It remains unclear whether the practice of flagging less trusted, or authoritative information, or offering alternative, more trusted or contextualised, related content is efficient.[165] As such, these initiatives could be further improved. In particular, they should, *inter alia,* include a charter of ethics comparable to the highest professional standards of journalism, be as open and transparent as possible and involve a wide range of stakeholders in order to ensure that Internet users receive a real diversity of opinions and ideas and are able to better identify misinformation.[166]

By contrast, initiatives such as Facebook's decision to prioritise content posted by friends at the expense of 'public' content seem to suggest that companies may prefer to evade responsibility for the quality of the information on its networks, rather than engage with content publishers. It is unlikely that this type of initiative would contribute to greater access to better quality information, and such initiatives are likely to run into difficulties when information news sources trusted by the community clash with information provided by governments.[167]

More concerning, however, is the lack of transparency surrounding the tweaking of companies' algorithms in order to produce reliable results or good content. In particular, there is a real risk that the content of small media companies might become less visible as a result, by pushing them further down the list of recommended content.[168] This raises significant issues for media pluralism, diversity and competition. A lack of transparency regarding the criteria and systems used by internet companies in order to identify and suspend so-called "fake" accounts on the basis of suspicious activity raises questions about the more detailed criteria and systems used by Internet companies in order to take such decisions. There is currently little to no information available to ensure that companies do not close accounts by mistake and what redress is available when errors are made.

Other restrictions

Real-name policies

While Twitter and YouTube do not have real-name policies, Facebook requires its users to use their real name when posting content on its platform. In ARTICLE 19's view, this is inconsistent with international standards on free expression and privacy, in particular:

- The use of real-name registration as a prerequisite to using their services can have a negative impact on users' rights to privacy and freedom of expression, particularly for those from minority groups, or those in a situation of heightened risk or vulnerability, who might be prevented from asserting their sense of identity.

- Whilst real-name policies are usually presented as an effective tool against internet trolling, fostering a culture of mutual respect between internet users, the disadvantages of real-name policies outweigh their benefits. In particular, anonymity is vital to protect children, victims of crime, individuals from minority groups and others in a situation of heightened risk or vulnerability from being targeted by criminals or other malevolent third parties who may abuse real-name policies. In this sense, anonymity is as much about online safety as self-expression.

Identification policies

ARTICLE 19 is further concerned that real-name policies are often accompanied by a requirement to provide identification. For instance, Facebook lists the different types of IDs it accepts in order to confirm its users' identity.[169] This, in our view, raises serious concerns over data protection, given that many such demands require users to provide a considerable amount of sensitive personal data as a means to verify their identity. Even if Facebook were to delete such data immediately, the very existence of the company's policy could put users at risk in certain countries. In particular, governments could more easily track down dissidents since they would already be identified by their Facebook account.

Content removal processes

ARTICLE 19 notes that while the mechanisms put in place by dominant social media platforms to remove content generally feature some procedural safeguards, none contain all the appropriate safeguards. As such, they all fall short of international standards on free expression and due process in some way.

Youtube

YouTube increasingly uses machine learning and algorithms to flag certain categories of content for removal;[170] such as 'extremist' content and copyright material.[171] Users

may also either flag videos[172] or file more detailed reports in relation to multiple videos, comments, or a user's entire account.[173] Different report forms are available depending on the type of complaint at issue (e.g. privacy or legal reporting, reporting on critical injury footage).[174] YouTube also relies on a trusted flagger system, whereby reports filed by trusted flaggers are fast-tracked for review. While it does not seem to enable a mediation process by way of counter-notice before material is taken down or other sanction is applied, it provides avenues of redress in relation to copyright,[175] account termination[176] and video strikes.[177] It is unclear, however, whether individuals are notified of the reasons for any measure taken by YouTube. In ARTICLE 19's view, this – and the lack of counter-notice mechanism before action – is the most significant shortfall of YouTube's internal review mechanism, which is otherwise broadly consistent with international standards on free expression and due process safeguards.

Facebook

Facebook provides various reporting mechanisms, from reporting particular accounts, pages or posts,[178] to "social reporting."[179] In addition, Facebook relies on algorithms to filter out certain types of content and uses a trusted-flagger system to fast-track reports of violations of its Community Standards. Facebook also places significant emphasis on end-user tools to address abuse they encounter on the platform, such as hiding news feeds, blocking individuals or unfriending them.[180] Once reports are received, it appears that Facebook does not notify users of the reasons behind any restrictions it may subsequently apply to their accounts;[181] nor does it seem that Facebook provides for any clear appeals or review mechanism of its decisions.[182] While social reporting and Facebook's emphasis on other tools to prevent exposure to undesirable content are welcome, the absence of a clear appeals mechanism in relation to wrongful content removals or other sanctions is a fundamental flaw in its internal system. The failure to provide reasons for content restrictions imposed is also inconsistent with due process safeguards.

Twitter

Twitter also relies on filters to remove certain types of content on its own initiative. It also provides for different types of reporting mechanisms, depending on the nature of the complaint at issue.[183] Reporting forms are generally comprehensive.[184] By contrast, it is difficult to find information about any appeals mechanism to challenge a decision by Twitter to take action in relation to either an account or particular content. Rather than providing a separate page dealing with such mechanisms, a link is provided on a seemingly *ad hoc* basis at the end of some of its content-related policies.[185] It therefore appears that any individual whose content is removed on the basis of Twitter's policies is generally not given any reasons for the decision, or a clear opportunity to appeal. Appeals only seem to be available when accounts are suspended.[186]

Positively, it appears that Twitter generally makes good faith efforts to inform users about legal requests it receives to remove content.[187] However, generally speaking, reasons for actions taken by Twitter, or access to an appeals mechanism, do not appear to be given on a consistent or systematic basis. In our view, these are significant shortfalls in Twitter's internal processes and inconsistent with due process safeguards.

Google

Google provides an easily findable form for legal content removal requests.[188] A more detailed form is available for so-called "right to be forgotten" requests.[189] By contrast, it is harder to find forms for reporting violations of Google's Terms of Service, which may partly be explained by the fact that different forms might be available on the website of different Google products. It is as a result highly unclear what appeals mechanisms are available in order to challenge the sanctioning measures that Google might apply. In general, it appears that individuals affected by such measures are not informed of Google's decision to de-list their content. In short, Google's processes lack transparency and fail to provide due process safeguards.

More generally, ARTICLE 19 notes that social media companies increasingly rely on algorithms and various forms of trusted-flagger systems in their content removal processes, sometimes preventing content from being published in the first place. We believe that it is imperative that companies are more transparent in relation to their use of these tools. In particular:

- The lack of transparency in relation to the use of algorithms to detect particular types of content means that they are more likely to be prone to bias. It is also unclear how algorithms can be trained to take into account free speech concerns, or the context of the content, if at all.

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

Finally, ARTICLE 19 notes that social media companies are sometimes required to put in place national contact points based on a specific law[190] They also sometimes appear to have voluntarily agreed to such a system.[191] This is a matter of concern,

particularly in countries with governments with a poor record on the protection of freedom of expression. In particular, national points of contact may facilitate the removal of content that would be considered legitimate under international human rights law. As such, we believe that social media companies should refrain from voluntarily putting in place national contact points, particularly in those countries where the respective governments have a poor record on the protection of freedom of expression.

Sanctions for failure to comply with Terms of Service

ARTICLE 19 notes that most social media companies rightfully apply different types of sanctions against users who infringe their Terms of Service, alongside using other tools allowing users to contact the authors of a post to seek a peaceful resolution of a complaint. Sanctions usually range from a simple strike against their account or the disabling of certain features, to geo-blocking or the termination of their account. For instance:

- In its Terms of Use, **YouTube** provide that it reserves the right to remove content and/or terminate a user's access for uploading content in violation of its Terms.[192] Its Reporting Centre page contains additional information about its account termination policy[193] as well as its video-strike policy and appeals mechanism.[194]

- **Facebook** provides, in its Community Standards, that "the consequences of breaching our Community Standards vary depending on the severity of the breach and a person's history on Facebook. For instance, we may warn someone for a first breach, but if they continue to breach our policies, we may restrict their ability to post on Facebook or disable their profile."[195]

- **Twitter** explains in its Twitter Rules that "all individuals accessing or using Twitter's services must adhere to the policies set forth in the Twitter Rules. Failure to do so may result in Twitter taking one or more of the following enforcement actions: (i) requiring you to delete prohibited content before you can again create new posts and interact with other Twitter users; (ii) temporarily limiting your ability to create posts or interact with other Twitter users; (iii) asking you to verify account ownership with a phone number or email address; or (iv) permanently suspending your account(s).[196] Twitter also has a dedicated page explaining its enforcement options at various levels, from response to tweets to direct messages and accounts.[197] It also clearly explains the general principles guiding the enforcement of its policies.[198]

- **Google** does not provide clear guidance as to the options available to it when requests are made to de-index links from its search engines. In particular, Google does not explain whether links removed from its search engines are removed globally or on a country-level basis. Google has previously been

criticised for removing copyright infringing content globally, whilst refusing to do so for other types of content, such as privacy infringements.[199]

In ARTICLE 19's view, with the exception of Google whose policies in this area generally lack transparency, the above Terms are broadly consistent with international standards on freedom of expression and the *Manila Principles on Intermediary Liability*. These standards provide that content restriction policies and practices must comply with the tests of necessity and proportionality under human rights law. At the same time, we regret that these companies seem to be increasingly applying country filters so that the promise of free expression 'beyond borders' is rapidly evaporating.[200]

Types of regulation: policy options

Given the clear shortfalls of regulation by contract, various policy options are in existence (or are being considered) in this area.

Regulation

Although in a majority of countries, dominant social media companies have traditionally been regulated by way of conditional immunity from liability,[201] States are increasingly resorting to more intrusive forms of regulation.[202] In ARTICLE 19's view, many of these regulatory models are deeply problematic for several reasons:

- They tend to give disproportionate censorship powers to the State, whether through prison terms, fines or content blocking powers, chilling free expression. The underlying content laws that regulators are required to enforce are generally overly broad. In many countries, the regulator is not an independent body and the law does not always provide for a right of appeal or judicial review of the regulator's decisions.[203] Furthermore, by putting the State in the position of being the ultimate arbiter of what constitutes permissible expression or what measures companies should be adopting to tackle 'illegal' content, these regulatory models are more likely to undermine minority viewpoints.

- They are inconsistent with the international standards on freedom of expression outlined above, which mandate that approaches to regulation developed for other means of communication – such as telephony or broadcasting – cannot simply be transferred to the Internet but, rather, need to be specifically designed for it. [204]

- Sanctions powers including the blocking of entire platforms or hefty fines for failure to comply with domestic legal requirements would, in and of themselves, constitute a disproportionate restriction on freedom of expression.[205]

Co-regulation

Co-regulation – a regulatory regime involving private regulation that is actively encouraged or even supported by the State[206] – is also increasingly seen as an alternative to direct regulation or the mere application of companies' Terms of Service.[207] Co-regulation can include the recognition of self-regulatory bodies by public authorities. The latter generally also have the power to sanction any failure by self-regulatory bodies to perform the functions for which they were established.[208]

In ARTICLE 19's view, however, many types of co-regulatory models ultimately present the same flaws as regulation by entrusting too much power to state institutions to regulate online expression. This would not only have a chilling effect on freedom of expression but would also hamper innovation. By contrast, we note that simply providing legal underpinning to a self-regulatory body whilst at the same time guaranteeing the independence of such a body can be compatible with international standards on freedom of expression.[209]

Self-regulation

For the purposes of this policy brief, ARTICLE 19 considers regulation by contract (i.e. the application of its Terms of Service by a company) to be distinct from self-regulation. Self-regulation[210] is a framework that relies entirely on voluntary compliance: legislation plays no role in enforcing the relevant standards. Its *raison d'être* is holding members of self-regulatory bodies accountable to the public, promoting knowledge within its membership and developing and respecting ethical standards. Those who commit to self-regulation do so for positive reasons such as the desire to further the development and credibility of their sector. Self-regulation models rely first and foremost on members' common understanding of the values and ethics that underpin their professional conduct – usually in dedicated "codes of conduct" or ethical codes. Meanwhile, members seek to ensure that these voluntary codes correspond to their own internal practices.

Increasingly, companies are threatened with regulation if they fail to abide by the standards laid down in self-regulatory codes and, in some instances, specific codes of conduct have been adopted jointly by companies and public institutions.[211] However, these types of "voluntary" initiatives are often used to circumvent the rule of law. They lack transparency and fail to hold companies and governments accountable for wrongful removal of content.

ARTICLE 19's position

Rather than seeking to regulate or co-regulate, ARTICLE 19 believes that:

- **Companies should respect the Guiding Principles on Business and Human Rights**. Although companies are in principle free to set their own terms and conditions subject to well-established contract law exceptions (such as illegality), which may vitiate the validity of the contract, they should respect international standards on human rights consistent with the Guiding Principles.[212] This is especially important in the case of the dominant social media companies, which have now become central enablers of freedom of expression online. In its most basic sense, adherence to the Guiding Principles means that Community Standards should be in line with international standards on freedom of expression and that private companies should provide a remedy for free speech violations under their Community Standards. Moreover, as a matter of

constitutional theory, it is at least arguable that Community Standards should be applied in a manner consistent with constitutional values and standards that include the protection of freedom of expression.[213]

- **States must provide for an effective remedy for free speech violations by private parties**, particularly in circumstances where companies unduly interfere with individuals' right to freedom of expression by arbitrarily removing content or imposing other restrictions on freedom of speech. This approach is consistent with States' positive obligation to protect freedom of expression under international human rights law.[214] In practice, we believe that the creation of a new cause of action could be derived either from traditional tort law principles or, as noted above, the application of constitutional theory to the enforcement of contracts between private parties.[215] Moreover, and in any event, we note that private companies are already subject to laws that protect constitutional principles such as non-discrimination or data protection with horizontal effect. Accordingly, given that private companies are required to comply with basic constitutional values, there is no reason in principle why they should not also be required to comply with international human rights standards, including the right to freedom of expression.

- **Companies should explore the possibility of independent self-regulation.** ARTICLE 19 believes that the need for an effective remedy for free speech violations could also be addressed by companies collaborating with other stakeholders to develop new self-regulatory mechanisms **such as a 'social media council.'**[216] In our view, this model could provide an appropriate framework for addressing current problems with content moderation by social media companies, provided that it meets certain conditions of independence, openness to civil society participation, accountability and effectiveness. Independent self-regulation would also allow for the adoption of adapted and adaptable remedies unhindered by the threat of heavy legal sanctions. It would also foster greater transparency in companies' use of algorithms to distribute or otherwise control content. In sum, this mechanism could constitute a transparent and accountable forum for public debate on issues related to online circulation of content; it would also foster greater transparency in the use of algorithms to distribute content.

ARTICLE 19's recommendations

Recommendations to States

Recommendation 1: Intermediaries should be shielded from liability for third-party content

ARTICLE 19 reiterates that social media platforms should be immune from liability for third-party content in circumstances where they have not been involved in modifying the content at issue.[217] States should adopt laws to that effect and refrain from adopting laws that would make social media companies subject to broadcasting regulators or other similar public authorities. Equally, intermediary liability laws should be interpreted so that social media companies do not lose immunity from liability simply because they have content moderation policies in place and are in principle able to remove online content upon notice.

Recommendation 2: There should be no extra-legal pressure on intermediaries

ARTICLE 19 recommends that States should not use extra-judicial measures or seek to bypass democratic or other legal processes to restrict content. In particular, they should not promote or enforce so-called "voluntary" practices or secure agreements that would restrain public dissemination of content. At the same time, ARTICLE 19 notes that companies are under no obligation to comply with government requests that have no basis in law. They remain free to decide whether such requests are in breach of their Terms of Service.

Recommendation 3: Right to an effective remedy between private parties should be provided

ARTICLE 19 recommends that States should address a lack of remedy for the wrongful removal of content on the basis of companies' Terms of Service and redress this procedural asymmetry. States have an obligation to provide an effective remedy under international law, which, arguably, applies to interference by a private party with the free expression rights of another private party. In practice, this means that individuals should be provided with an avenue of appeal once they have exhausted social media companies' internal mechanisms. This could be an appeal to the courts or a consumer or other independent body.[218] In such cases, the courts could address themselves to the question of whether the social media platform in question had acted unfairly or unreasonably by removing the content at issue. This question could be decided in light of international standards on freedom of expression or equivalent constitutional principles.

Further, or in the alternative, States should seek to develop new causes of action, such as tortious interference with free expression rights. Basic elements could include any serious damage to the ability of individuals to share lawful information. While the exact contours of such a new tort are beyond the scope of the present paper, further research could be undertaken in this area.

Recommendations to social media companies

Recommendation 1: Terms of Service should be sufficiently clear, accessible and in line with international standards

Consistent with the Guiding Principles on Business and Human Rights, ARTICLE 19 believes that dominant social media companies have a responsibility to respect international standards on freedom of expression and privacy. In practice, this entails the following recommendations:

- At a minimum, companies' Terms of Service must be sufficiently clear and accessible so as to enable their users to know what is and what is not permitted on the platform, and regulate their conduct accordingly. In practice, this means, among other things, that companies should explain their Terms of Service in plain language, where possible, and translate them into the language of the countries in which they operate. Terms of Service should also be made available in different formats to facilitate access.

- Social media platforms should not require the use of real names. At the very least, Internet companies should ensure anonymity remains a genuine option for users. Equally, social media platforms should not require their users to identify themselves by means of a government-issued document or other form of identification.

- Terms of Service should comply with international standards on freedom of expression. In particular, social media companies should provide specific examples addressing the way in which their standards are applied in practice (such as through the use of case studies). This should be accompanied by guidance as to the factors that are taken into account in deciding whether or not content should be restricted.

- Companies should conduct regular reviews of their Terms of Service to ensure compliance with international standards on freedom of expression both in terms of their formulation and their application in practice. In particular, they should conduct regular audits or human rights impact assessments designed to monitor the extent to which content moderation policies adhere to the principle of non-discrimination. This would at least go some way towards guaranteeing the free expression rights of minority and marginalised groups. Users should be clearly notified of any changes in companies' Terms of Service as a result of such reviews or human rights impact assessments.

Recommendation 2: Content removal processes, including the use of algorithms and trusted-flagger schemes should be fully transparent

ARTICLE 19 urges social media companies to be transparent in relation to their use of algorithms and 'trusted-flagger' systems. In particular:

- Companies should provide sufficient information for the public to understand how algorithms operate to detect unlawful or harmful content. Given that algorithms currently have very limited ability to assess context, Internet intermediaries should at the very least ensure human review of content flagged through these mechanisms. More generally, Internet companies should conduct human rights impact assessments of their automated content management systems, and in particular the extent to which they are likely to lead to over-removal of (legitimate) content.

- Companies should clearly identify which government or other third-party organisations are given the status of trusted-flagger and explain what criteria are being used to grant this status.

More generally, before taking operational decisions that are likely to have a significant impact on human rights, companies should conduct a human rights impact assessment of the extent to which human rights are likely be negatively impacted. For example, before voluntarily putting in place national contact points, social media companies should consider the likely impact of such a decision on freedom of expression in the particular country. In countries with governments with a poor record on the protection of freedom of expression, social media companies should refrain from taking such measures.

Recommendation 3: Sanctions for failure to comply with Terms of Service should be proportionate

ARTICLE 19 further recommends that companies should ensure that sanctions for failure to comply with their Terms of Service are proportionate. In particular, we recommend that social media platforms should:

- Be clear and transparent about their sanctions policy; and

- Apply sanctions proportionately so that the least restrictive technical means should be adopted. In particular, the termination of an account should be a measure of last resort that should only be applied in the most exceptional and serious circumstances.

Moreover, and in any event, users should be encouraged in the first instance to use the technological tools available on the platform to block users or prevent certain types of content from appearing in their news feed where appropriate.

Recommendation 4: Internal complaints mechanisms should be provided

ARTICLE 19 recommends that in order to respect international standards on freedom of expression, social media companies should put in place effective complaints mechanisms to remedy the wrongful removal of content, or other disproportionate restrictions on the exercise of their users' right to freedom of expression. In doing so, they should respect basic due process rights.

In particular, we believe that individuals should be given notice that a complaint has been made about their content. They should also be given an opportunity to respond before the content is taken down or any other measure is applied by the intermediary. In order for individuals to be able to respond effectively, the notice of complaint should be sufficiently detailed.[219] If the company concludes that the content should be removed or other restrictive measures should be applied, individuals should be notified of the reasons for the decision and given a right to appeal the decision.

In circumstances where the company has put in place an internal mechanism, whereby it takes down content merely upon notice, we believe that at a minimum, the intermediary should:

- Require the complainant to fill out a detailed notice,[220] in which they identify the content at issue; explain their grounds for seeking the removal of content; and provide their contact details and a declaration of good faith.
- Notify the content producer that their content has been removed or that another measure has been applied to their account.
- Give reasons for the decision,
- Provide and explain internal avenues of appeal.

ARTICLE 19 also recommends that companies should also ensure that appeals mechanisms are clearly accessible and easy to find on their platform. As platforms that rely on their users' free expression as their business model, they should ensure that the right to freedom of expression is protected through adequate procedural safeguards on their platform.

Recommendation 5: Independent self-regulatory mechanism should be explored to provide greater accountability

In addition to internal complaints mechanisms, ARTICLE 19 recommends that social media companies collaborate with other stakeholders to develop new independent self-regulatory mechanisms for social media. This could include a dedicated social media council – inspired by the effective self-regulation models created to promote journalistic ethics and high standards in print media. Such a mechanism should be established preferably at national level with some international coordination. It would involve the development of an Ethics Charter for social media platforms

and the digital distribution of content and would promote the use of non-pecuniary remedies in order to deal with wrongful removal of content. It would also foster greater transparency in the use of algorithms to distribute or otherwise control content.

Recommendation 6: Government and court orders in breach of international human standards should be resisted

ARTICLE 19 further believes that for companies to demonstrate their commitment to respect human rights, they should challenge government and court orders that they consider to be in breach of international standards on freedom of expression and privacy. In practice, this means that as a matter of principle, companies should:

- Resist government legal requests to restrict content in circumstances where they believe that the request lacks a legal basis or is disproportionate. This includes challenging such orders before the courts.
- Resist individuals' legal requests to remove content in circumstances where they believe that the request lacks a legal basis or is disproportionate.
- Appeal court orders demanding the restriction of access to content that is legitimate under international human rights law.

Moreover, as a matter of principle, companies should resist government extra-legal or informal requests to restrict content on the basis of their Terms of Service. In this regard, companies should make clear to both governments and private parties that they are under no obligation to remove content when a takedown request is made on the basis of their Terms of Service.

The same principles apply to government requests – whether legal or informal – to provide user data.[221] This is particularly important when the user in question is a human rights defender, protester or other dissenter. In this regard, we note that if companies were to voluntarily share their users' data with governments that are known for cracking down on dissent, they would arguably be complicit in human rights violations.

Recommendation 7: Transparency reporting must be more comprehensive

ARTICLE 19 believes that companies' transparency reporting, whilst positive, should be further improved. In particular, social companies should specify whenever they remove content on the basis of their Terms of Service at the request of governments or 'trusted' third parties, such as local NGOs or associations. Moreover, companies should provide information about the number of complaints they receive about alleged wrongful removals of content and the outcome of such complaints (i.e. whether content was restored or not). More generally, we recommend that transparency reports should contain the types of information listed in the Ranking Digital Rights indicators for freedom of expression.[222]

Endnotes

[1] Facebook Stats, which lists 1.45 billion daily active users on average for March 2018 and 2.20 billion monthly active users as of 31 March 2018.

[2] Statistics Portal, Number of monthly active Twitter users worldwide from 1st quarter 2010 to 1st quarter 2018 (in millions).

[3] YouTube for Press, YouTube in numbers.

[4] ARTICLE 19, Internet Intermediaries: Dilemma of Liability, 2013.

[5] The analysis of the Terms of Service of selected social media companies is published separately; forthcoming, available on ARTICLE 19's website.

[6] ARTICLE 19, Getting connected: Freedom of expression, telcos and ISPs, 2017.

[7] Through its adoption in a resolution of the UN General Assembly, the UDHR is not strictly binding on states. However, many of its provisions are regarded as having acquired legal force as customary international law since its adoption in 1948; see *Filartiga v. Pena-Irala*, 630 F. 2d 876 (1980) (US Circuit Court of Appeals, 2nd circuit).

[8] UN General Assembly, International Covenant on Civil and Political Rights, 16 December 1966, UN Treaty Series, vol. 999, p. 171.

[9] Article 10 of the European Convention for the Protection of Human Rights and Fundamental Freedoms, 4 September 1950; Article 9 of the African Charter on Human and Peoples' Rights (Banjul Charter), 27 June 1981; Article 13 of the American Convention on Human Rights, 22 November 1969.

[10] HR Committee, General Comment No. 34 on Article 19: Freedoms of opinion and expression, CCPR/C/GC/34, 12 September 2011, para 12, 17 and 39.

[11] The 2011 Joint Declaration on Freedom of Expression and the Internet, 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, June 2011.

[12] *Ibid*.

[13] *Ibid*. See also the Report of the Special Rapporteur on FOE, A/66/290, 10 August 2011, para 16.

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

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

[16] The 2011 Joint Declaration, *op. cit*.

[17] The Report of the Special Rapporteur on FOE, 16 May 2011, A/HRC/17/27, para 43.

[18] *Ibid.,* para 42.

[19] 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.

[20] *Ibid.*, Principle 15.

[21] The May 2011 Report of the Special Rapporteur on FOE, *op.cit.*, para 45.

[22] *Ibid.* para 46.

[23] *Ibid.*, paras 47 and 76.

[24] *Ibid.*, paras 48 and 77.

[25] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, 11 May 2016, A/HRC/32/38; para 40 – 44.

[26] *Ibid.*

[27] *Ibid.*, para 43.

[28] OAS Special Rapporteur on FOE, Freedom of Expression and the Internet, 2013. The report noted that "the adoption of voluntary measures by intermediaries that restrict the freedom of expression of the users of their services - for example, by moderating user-generated content - can only be considered legitimate when those restrictions do not arbitrarily hinder or impede a person's opportunity for expression on the Internet;" paras 110-116.

[29] *Ibid.*, paras 111-112.

[30] *Ibid.*, para 113.

[31] *Ibid.*, para 114.

[32] *Ibid.*, para 115.

[33] *Ibid.*, para 116.

[34] OAS Special Rapporteur on FOE, Standards for a Free, Open and Inclusive Internet, 2016, paras 95-101.

[35] *Ibid.*, para 98.

[36] *Ibid.*, para 99.

[37] 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.

[38] *Ibid.*, p. 24.

[39] Committee of Ministers of Council of Europe, Recommendation CM/Rec (2012)4 of the Committee of Ministers to Member States on the protection of human rights with regard to social networking services, *adopted by the Committee of Ministers on 4 April 2012 at the 1139th meeting of the Ministers' Deputies.* These recommendations were further echoed in the Committee of Ministers Guide to Human Rights for Internet users, which states "your Internet service provider and your provider of online content and services have corporate responsibilities to respect your human rights and provide mechanisms to respond to your claims. You should be aware, however, that online service providers, such as social networks, may restrict certain types of content and behaviour due to their content policies. You should be informed of possible restrictions so that you are able to take an informed decision as to whether to use the service or not. This includes specific information on what the online service provider considers as illegal or inappropriate content and behaviour when using the service and how it is dealt with by the provider;" Guide to human rights for Internet users, Recommendation CM/Rec(2014)6 and explanatory memorandum, p. 4.

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

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

[42] *Ibid.,* Principle IV.

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

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

[45] 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.

[46] 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).

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

[48] The Joint Declaration on Freedom of Expression and "Fake News", Disinformation and Propaganda, adopted by the UN Special Rapporteur on FOE, the OSCE Representative on Freedom of the Media, the OAS Special Rapporteur on Freedom of Expression and the ACHPR Special Rapporteur on Freedom of Expression and Access to Information, 3 March 2017, para 4 a).

[49] *Ibid.*

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

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

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

[53] *Ibid.*

[54] The right of private communications is protected in international law through Article 17 of the ICCPR, *op.cit.*, which provides, *inter alia*, that: "No one shall be subjected to arbitrary or unlawful interference with his privacy, family, home or correspondence, nor to unlawful attacks on his honour and reputation." The UN Special Rapporteur on promotion and protection of human rights and fundamental freedoms while countering terrorism has argued that like restrictions on the right to freedom of expression under Article 19, restrictions of the right to privacy under Article 17 of the ICCPR should be interpreted as subject to the three-part test; see the Report of the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, Martin Scheinin, A/HRC/13/37, 28 December 2009.

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

[56] *Ibid.*, para 84.

[57] Report of the Special Rapporteur to the Human Rights Council on the use of encryption and anonymity to exercise the rights to freedom of opinion and expression in the digital age, A/HRC/29/32, 22 May 2015, para 60.

[58] *Ibid.*

[59] This is also consistent with Phase I, Pilot Study of the Ranking Digital Rights Project, March 2015, p. 15. For example, Twitter now publishes information about government removal requests based on the company's Terms of Service, see Twitter, Government TOS reports; Google has overhauled its Transparency Report, which provides additional contextual information or data e.g. relating to traffic disruption or National Security Letters, see Google, A new look for our Transparency Report, 18 July 2017.

[60] Facebook's Transparency Report only records government requests. The report was last updated in December 2016.

[61] Google Transparency Report; Google are more explicit about the fact that their statistics do not cover removals on the basis of their terms of service when the request is made by third parties. At the same time, the Google Transparency Report explains that government requests' statistics includes requests made on the basis of its Terms of Use but it is not disaggregated from requests made on the basis of court orders or other legal authority.

[62] For example, YouTube states that videos are not automatically taken down by their community-driven flagging system; see Flagging Content on YouTube.

[63] Though good progress has been made in this area by YouTube in their new Community Guidelines Enforcement (as of May 2018).

[64] See, e.g., Wired, Google's Using a Combination of AI and Humans to Remove Extremist Videos on YouTube, 19 June 2017.

[65] See, e.g., Guardian, Facebook apologises for deleting free speech group's post on Syrian torture, 6 July 2012.

[66] See, e.g., Forbes, How Feminism Beat Facebook (And What The Campaign Might Mean For Online Equality), 30 May 2013. In that case, feminist groups believed that Facebook community guidelines were overly broad and campaigned for Facebook to review its policy on violent speech and images directed at women.

[67] For example, on or about January 2018, Facebook's Help Centre stated, "When something gets reported to Facebook, we'll review it and remove anything that doesn't follow the Facebook Community Standards. Your name and other personal information will be kept completely confidential if we reach out to the person responsible" (emphasis added). Facebook went on to explain that reporting content did not guarantee its removal but stayed silent on the review process. In April 2018, Facebook announced that it would be strengthening its appeals process in a post in its newsroom. However, it has yet to be rolled out and does not include a separate section on its website explaining the process. Rather it is premised on notification to the person whose content has been removed and offering the opportunity to request a review at the touch of a button.

[68] YouTube, Dispute A Content ID Claim. It is worth noting that this internal mechanism may to some extent be the result of the Digital Copyright Millennium Act (DMCA).

[69] Twitter, Legal request FAQs.

[70] *Ibid.*

[71] See, e.g., YouTube, Terms of Service (as of May 2018), which state "YouTube reserves the right to decide whether Content violates these Terms of Service for reasons other than copyright infringement, such as, but not limited to, pornography, obscenity, or excessive length. YouTube may at any time, without prior notice and in its sole discretion, remove such Content and/or terminate a user's account for submitting such material in violation of these Terms of Service."

[72] For example, the YouTube Terms of Service, *op.cit.,* stipulate that any contractual claim or dispute arising from the Service "shall be decided exclusively by a court of competent jurisdiction located in Santa Clara County, California." Confusingly, YouTube seems to publish a different version of its of Terms of Use, depending on the geolocation of the user. For instance, the UK version of its Terms of Service has a jurisdiction clause stating that the courts of England having exclusive jurisdiction to resolve any dispute arising from these Terms. Notwithstanding this, YouTube asserts its right to apply for injunctive remedies in any jurisdiction.

[73] This was confirmed in a relatively recent case, where a French tribunal found that a similar clause in Facebook Terms' of Use was unfair. The tribunal concluded that it had jurisdiction to hear a complaint about the termination of a Facebook user's account allegedly in violation of his free speech rights; see, Les Echos, Facebook will be prosecuted for censoring the Origin of the World and deletion of a profile, 05 March 2015. The European Union has recently stepped up pressure on Facebook, Google and Twitter to provide accessible and effective remedies, see Reuters, EU increases pressure on Facebook, Google and Twitter over user terms, 24 July 2017.

[74] EU, Facebook, Google and Twitter accept to change their terms of service to make them customer-friendly and compliant with EU rules, 15 February 2018.

[75] See, e.g., Eric Goldman, *Online User Account Termination and 47 U.S.C. §230 (c) (2)*, UC Irvine Law Review, Vol. 2, 2012.

[76] See, e.g., *Young v. Facebook*, 2010 WL 4269304 (N.D. Cal. Oct. 25, 2010) and related case comment.

[77] See, e.g., Le Figaro, The Origin of the World: Justice rejects the user who considered himself to be censored by Facebook,16 March 2018; by contrast, a German court recently issued an injunction preventing Facebook from removing a user's post. However, the reasons for the decision have apparently not been disclosed: see Bloomberg, Facebook told to stop deleting German user's immigrant comment, 12 April 2018.

[78] For example, there has been a wide criticism of the gendered approach of Facebook to nudity, which bans images of women's nipples and, up until relatively recently,

images of breast-feeding mothers. See, e.g., Gawker, New Yorker Temporarily Banned for From Facebook For Posting Cartoon Boobs, 9 October 2012. For Facebook's policy on breastfeeding, see here.

[79] Twitter Abusive Behaviour Policy; or Policy and product updates aimed at combatting abuse, April 2015.

[80] Facebook Community Standards (as of January 2018). Since then, Facebook has updated its Community Guidelines. They no longer refer to the "condoning" of terrorist or other violent activities. Although these policy updates may well reflect previously existing practices within these companies, they also appear to have been partly influenced by the coming into force of a French Decree allowing the administrative blocking of terrorist content - which, under French law, includes statements publicly *condoning* acts of terrorism online; see, e.g. Facebook, Explaining our Community Standards and Approach to Government Requests, 15 March 2015. See also Riga Joint Statement of Justice and Home Affairs Ministers, 29-30 January 2015 and Guardian, Facebook clarifies policy on nudity, hate speech and other community standards, 16 March 2015.

[81] See, e.g., Business Insider, Facebook Will Block Photos Celebrating Rape Following Ad Boycott, 28 May 2013

[82] Facebook Community Standards, *op.cit.*

[83] See, e.g., Financial Times, Google apologises to advertisers for extremist content on YouTube, 20 March 2017.

[84] See, e.g., Financial Times, Four ways Google will help to tackle extremism, 18 June 2017.

[85] See, e.g., EDRI, Human Rights and Privatised Law Enforcement, 25 February 2014

[86] This a particular issue with the Anti-Counterfeiting Trade Agreement; see, e.g., EDRI, The ACTA archive, 04 July 2013.

[87] While 'hate speech' has no definition under international human rights law, 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: a) severe forms of 'hate speech' that international law *requires* States to prohibit; b) other forms of 'hate speech' that States *may* prohibit; and c) 'hate speech' that is lawful but nevertheless raises concerns in terms of intolerance and discrimination, meriting a critical response by the State, but should be protected. Given the complexity around this term, ARTICLE 19 refers to it as *'hate speech'* in this and other policy documents. For an overview of ARTICLE 19 polices applicable to 'hate speech,' see ARTICLE 19, Hate Speech Explained: a Toolkit, 2015.

[88] While 'hate speech' has no definition under international human rights law, 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: a) severe forms of 'hate speech' that international law requires States to prohibit; b) other forms of 'hate speech' that States may prohibit; and c) 'hate speech' that is lawful but nevertheless raises concerns in terms of intolerance and discrimination, meriting a critical response by the State, but should be protected. Given the complexity around this term, ARTICLE 19 refers to it as 'hate speech' in this and other policy documents.

[89] See, e.g., Le Monde, Associations will sue the three giants of the American Internet, 15 May 2016.

[90] Facebook updated its community standards on 'hate speech' (as a part of 'objectionable content') in 2015 (*op. cit.*) and in 2018 in an attempt to clarify the criteria that the company uses to remove content.

[91] Facebook, Objectionable Content, Hate Speech (as of May 2018).

[92] *Ibid.*

[93] *Ibid.*

[94] *Ibid.* These include expression such as e.g. "ugly," "hideous," "deformed," "retarded," "stupid," "idiot," "slutty," "cheap," "free-riders," or statement of contempt, such as "I hate," "I don't like," "X are the worst;" or expressions of disgust, such as "vile," "gross," or "disgusting."

[95] *Ibid.* Slurs are defined as words that are commonly used as insulting labels for the above-listed characteristics.

[96] Facebook, Community Standards, Dangerous Individuals and Organizations. Hate organisations are defined as "any association of three or more people that is organised under a name, sign or symbol and which has an ideology, statements or physical actions that attack individuals based on characteristics, including race, religious affiliation, nationality, ethnicity, gender, sex, sexual orientation, serious disease or disability."

[97] *Ibid.*

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

[99] Twitter, The Twitter Rules.

[100] Twitter, Hateful Conduct Policy.

[101] Though parody is the subject of a dedicated, see Twitter, Parody, newsfeed, commentary, and fan account policy.

[102] Twitter, Abusive profile information.

[103] YouTube, Hate Speech Policy.

[104] *Ibid.*

[105] *Ibid.*

[106] *Ibid.*

[107] Google, User Content and Conduct Policy.

[108] Facebook, Dangerous Organisations, *op. cit.*

[109] Ibid.

[110] See, e.g., the Report of the UN Special Rapporteur on Counter-Terrorism, A/HRC/16/51, 22 December 2010.

[111] *Ibid.* 'Terrorist organisations' are defined as "any non-governmental organisation that engages in premeditated acts of violence against persons or property to intimidate a civilian population, government or international organisation in order to achieve a political, religious or ideological aim" and "terrorist" as "a member of a terrorist organisation or any person who commits a terrorist act is considered a terrorist." A terrorist act is defined as "a premeditated act of violence against persons or property carried out by a non-government actor to intimidate a civilian population, government or international organisation in order to achieve a political, religious or ideological aim."

[112] *Ibid.*

[113] See, e.g., Dhaka Tribune, Facebook brands ARSA a dangerous organization, bans posts, 20 September 2017.

[114] A/66/290, *op. cit.* See also, ARTICLE 19, The Johannesburg Principles on National Security, Freedom of Expression and Access to Information, 1996.

[115] The August 2011 Report of the Special Rapporteur on FOE, *op. cit.*

[116] Facebook, Community Standards, Graphic Violence.

[117] This includes, among other things, imagery featuring mutilated people in a medical setting; videos of self-immolation when that action is a form of political speech or newsworthy; photos of wounded or dead people; videos of child or animal abuse; and videos that show acts of torture committed against a person or people.

[118] Twitter Rules, *op.cit.*

[119] Twitter, Violent extremist groups.

[120] *Ibid.*

[121] Twitter, Violent threats and glorification of violence.

[122] *Ibid.* "Glorification of violence" is defined as "behaviour that condones or celebrates violence (and/or its perpetrators) in a manner that may promote imitation of the act" or "where protected categories have been the primary target or victim."

[123] *Op. cit.*

[124] *Ibid.*

[125] See, e.g., European Court, *Gözel et Özer v. Turkey*, App. Nos. 43453/04 & 31098/05, 6 July 2010; and *Nedim Şener v. Turkey*, App.

No. 38270/11, para 115, 8 July 2014.

[126] YouTube policy, *op. cit.*

[127] *Ibid.*

[128] *Op. cit.*

[129] *Ibid.*

[130] The Twitter Rules state that users "may not engage in the targeted harassment of someone, or incite other people to do so;" the Rules also explain that abusive behaviour consists in "an attempt to harass, intimidate, or silence someone else's voice."

[131] YouTube does not appear to have a dedicated policy to deal with harassment or cyberbullying. However, it makes clear that it will remove threats of serious physical harm against a specific individual or defined group of individuals; see YouTube, Policy on Threats.

[132] For example, Facebook restricts this content under the sections 'Safety,' '**Cruel and Insensitive**,' 'Objectionable content,' 'Credible Threats Violence' and 'Violence and Criminal Behaviour' of its Community Standards, *op.cit.* For example:

- Bullying content is removed insofar as it "purposefully targets private individuals with the intention of degrading or shaming them. Users cannot post *inter alia* the following types of content: (i) content about another private individual that reflects claims about sexual activity, degrading physical descriptions about or ranking individuals on physical appearance or personality, threats of sexual touching, sexualised text targeting another individual or physical bullying where context further degrades the individual (ii) content that has been "photoshopped" to target and demean an individual, including by highlighting specific physical characteristics or threatening violence in text or with imagery; and (iii) content that specifies an individual as the target of statements of intent to commit violence, calls for action of violence, statements advocating violence, aspirational and conditional statements of violence, "physical bullying," Facebook may also remove Pages or Groups that are dedicated to attacking individuals by, e.g. cursing, negative character claims or negative ability claims. When minors are involved, this content is removed, alongside other types of content, such as attacks on minors by negative physical description.

- Harassment is not expressly defined, thought the Community Standards provide examples of prohibited content. Contrary to its bullying policy, they clearly state that context and intent matter, and that it allows people to share and re-share posts "if it is clear that something was shared in order to condemn or draw attention to harassment."

- 'Cruel and insensitive content' is defined as a content that targets victims of serious physical or emotional harm." As such users should not post content that depicts real people and "mocks their implied or actual serious physical injuries, disease or disability, non-consensual sexual touching or premature death.

- As for 'credible threats of violence,' Facebook considers factors such as "language, context and details in order to distinguish casual statements from content that constitutes a credible threat to public or personal safety;" as well as "person's public visibility and vulnerability," and any additional information such as threats mentioning a target and a bounty/demand for payment, a reference or image of a specific weapon, details about location, timing and method etc.

[133] The Community Standards explain that Facebook restricts the display of nudity or sexual activity "because some people in our community may be sensitive to this type of content" and it "default[s] to removing sexual imagery to prevent the sharing of non-consensual or underage content" and

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

[134] The application of these rules still raise freedom of expression issues in practice, for instance with the wrongful removal of a famous photograph of a 9-year old naked girl fleeing napalm bombings in Vietnam: see ARTICLE 19, Facebook v Norway: learning how to protect freedom of expression in the face of social media giants, 14 September 2016.

[135] YouTube, Nudity and sexual content policies explains that "in most cases, violent, graphic, or humiliating fetishes are not allowed" but 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." YouTube encourages users to provide context the title and description of a video in order to determine the primary purpose of the video

[136] These include i.e. the sending of unwanted sexual content, objectifying the recipient in a sexually explicit manner, or otherwise engaging in sexual misconduct.

[137] Twitter, Twitter Media Policy.

[138] Google has a number of policies in place to deal with content that might infringe privacy. It may remove personal information, including National identification numbers like U.S. Social Security Number, Argentine Single Tax Identification Number, Brazil Cadastro de pessoas Físicas, Korea Resident Registration Number, China Resident Identity Card, and similarly; bank account numbers; credit card numbers; images of signatures; nude or sexually explicit images that were uploaded or shared without your consent and confidential, personal medical records of private people. Google further explains that it does not usually remove dates of birth, addresses and telephone numbers

[139] Twitter has dedicated policies to deal with "personal information", "intimate media," and media content. In addition, Twitter's rules on abusive behaviour, including "abuse" and "unwanted sexual advances" are also relevant to the protection of privacy. The policy on private information prohibits the publication of other people's private information without their express authorization and permission. Definitions of private information may vary depending on local laws but may include private contact or financial information, such as credit card information, social security or other national identity number, addresses or locations that are considered and treated as private, non-public, personal phone numbers, non-public, personal email addresses. Twitter also has a dedicated policy to deal with the sharing of intimate photos or videos of someone that were produced or distributed without their consent. It provides a non-exhaustive list of examples of intimate media that violate their policy, including hidden camera content involving nudity, partial nudity, and/or sexual acts; images or videos that appear to have been taken secretly and in a way that allows the user to see the other person's genitals, buttocks, or breasts (content sometimes referred to "creepshots" or "upskirts"); images or videos captured in a private setting and not intended for public distribution; and images or videos that are considered and treated as private under applicable laws. Meanwhile, Twitter deals with adult content under its media policy, which allows some forms of adult content in Tweets marked as containing sensitive media. However, such content may not be used in people's profiles or header images.

[140] Google, European privacy requests Search removals FAQs. The policy and processes on "right to be forgotten" were developed fol-

lowing a landmark decision of the European Court of Justice in this area.

[141] Facebook Community Standards, Spam.

[142] Facebook Community Standards, Misrepresentation. Facebook justifies these rules and its real-name policy by reference to the need for trust and accountability on its platform.

[143] See, e.g. The Guardian, Facebook to begin flagging fake news in response to mounting criticism, 15 December 2016.

[144] Ibid.

[145] See, e.g. The Guardian, Facebook promised to tackle fake news. But the evidence shows it's not working, 16 May 2017.

[146] Facebook Newsroom, Replacing Disputed Flags with Related Articles, 20 December 2017.

[147] Facebook Newsroom, Bringing People Closer Together, January 11, 2018.

[148] Facebook Newsroom, Helping Ensure News on Facebook Is From Trusted Sources, 19 January 2018.

[149] *Ibid.*

[150] Facebook Community Standards, Fake News.

[151] *Ibid.*

[152] Facebook Newsroom, Update on German Elections, 27 September 2017.

[153] Facebook Newsroom, Improving Enforcement and Transparency of Ads on Facebook, 2 October 2017.

[154] Facebook Newsroom, New Test to Provide Context About Articles, 5 October 2017.

[155] YouTube, Spam, deceptive practices & scams policies.

[156] BBC, YouTube to offer fake news workshops to teenagers, 21 Apr 2017.

[157] Market Watch, YouTube cracks down on conspiracies, fake news, 5 October 2017.

[158] Twitter, Impersonation Policy.

[159] Twitter, About specific instances when a Tweet's reach may be limited.

[160] Twitter, Our Approach to Bots & Misinformation, 14 June 2017.

[161] Twitter Public Policy, Update: Russian Interference in 2016 US Election, Bots, & Misinformation, 28 September 2017.

[162] New York Times, In Race Against Fake News, Google and Facebook Stroll to the Starting Line, 25 January 2017.

[163] Google, How we fought bad ads, sites and scammers in 2016, 25 January 2017.

[164] Google News Initiative, Building a stronger future for journalism.

[165] See, e.g., Poynter, It's been a year since Facebook partnered with fact-checkers. How's it going?, 15 December 2017.

[166] See, e.g., ARTICLE 19, Social media and fake news from a free speech perspective, 25 November 2016; for more details, see also ARTICLE 19's policy on fake news (forthcoming).

[167] New York Times, Facebook to Let Users Rank Credibility of News, 19 January 2018.

[168] For the information on the impact of Facebook's changes to its news feed in some test countries, including Cambodia, see, e.g., Forbes, Facebook's Explore Feed Experiment Is Already Crushing Cambodia's Businesses, 2 November 2017.

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

[170] See, e.g., Wired, Google's using a combination of AI and humans to remove extremist videos from YouTube, 19 June 2017.

[171] ARTICLE 19, Algorithms and automated decision-making in the context of crime prevention, 2 December 2016.

[172] YouTube, Report inappropriate content.

[173] *Ibid.*

[174] *Ibid.*

[175] YouTube, Counter Notification Basics.

[176] YouTube, Account terminations.

[177] *Ibid.*

[178] Facebook Help Centre, How to Report Things.

[179] Facebook Help Centre, What is social reporting? Social reporting are communication tools that allow users to request other users

to take content down.

[180] Facebook Help Centre, How do I report inappropriate or abusive things on Facebook (ex: nudity, hate speech, threats)?

[181] Facebook Help Centre, Report Something.

[182] See, however, Facebook's April 2018 announcement that it would be strengthening its appeals process, see Facebook Newsroom, Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process, 24 April 2018.

[183] Twitter Help Centre, Report violations.

[184] Twitter Help Centre, Report abusive behavior.

[185] Violent extremist groups, op.cit. and Twitter Help Centre, Appeal an account suspension or locked account.

[186] *Ibid.*

[187] Twitter Help Centre, Content removal requests.

[188] Google, Legal Removal Requests.

[189] Google, EU Privacy Removal, Personal Information Removal Request Form.

[190] See, e.g., Germany, see ARTICLE 19, Germany: The Act to Improve Enforcement of the Law in Social Networks, August 2017.

[191] New York Times, Facebook Faces a New World as Officials Rein In a Wild Web, 17 September 2017.

[192] YouTube, Account terminations.

[193] *Ibid.*

[194] YouTube, Appeal Community Guidelines actions.

[195] Facebook Community Standards, *op.cit.*

[196] Twitter Rules, *op.cit.*

[197] Twitter Help Centre, Our range of enforcement options.

[198] Twitter Help Centre, Our approach to policy development and enforcement philosophy.

[199] See, e.g., EDRI, Google's forgetful approach to the "right to be forgotten," 14 December 2016.

[200] Twitter Help Centre, About country withheld content.

[201] See, e.g., Internet intermediaries: Dilemma of Liability, *op. cit.*; or Stanford's World Intermediary Liability Map.

[202] For example in Russia, the government media watchdog, *Roskomnadzor*, can impose fines, suspend the operations of, or block access to social media platforms that fail to remove extremist content. Up until recently, social media platforms were required to register with *Roskomnadzor*. This system was recently repealed by Federal Law № 276-FZ. For more information, see, e.g., ARTICLE 19, Russia: Changes in the Sphere of Media and Internet Regulation, March 2016; Guardian, Russia threatens to ban Google, Twitter and Facebook over extremist content, 20 May 2015. In France, failure to takedown content inciting or promoting terrorism upon notice by law enforcement agencies can result in social media platforms being blocked. Also, in 2017, Germany adopted a law requiring social networks to adopt complaints procedures to deal with illegal content under the German Criminal Code, to delete or block access to that content, and to report periodically on the application of these measures. Failure to abide by these requirements is punishable with severe administrative fines of up to 50 million euros; for more details, see ARTICLE 19 analysis of the Law, *op.cit.*

[203] See, e.g., Tanzania Electronic and Postal Communications (Online content) Regulations 2018, 16 May 2018.

[204] The 2011 Joint Declaration on Freedom of Expression, *op.cit.*

[205] See, e.g., European Court, *Yildirim v Turkey,* App. no. 3111/10, 18 December 2012.

[206] See, e.g., C.T. Marsden, Internet Co-Regulation: European Law, Regulatory Governance, and Legitimacy in Cyberspace. Cambridge, Cambridge University Press, 2011 p. 54. It defines co-regulation as "a regulatory regime involving private regulation that is actively encouraged or even supported by the state through legislation, funding, or other means of state support or institutional participation."

[207] For example, according to the EU, "co-regulation gives, in its minimal form, a legal link

between self-regulation and the national legislator in accordance with the legal traditions of the Member States. Co-regulation should allow for the possibility of State intervention in the event of its objectives not being met;" see Recital 44 of the Audio-Visual Media Services Directive 2010/13/EU.

[208] Co-regulation of this kind has recently emerged in Germany, *op.cit.* Under the 2017 law, government authorities can "recognise" – or withdraw recognition of - self-regulatory bodies, whose purpose is to supervise content moderation on social media platforms. In other words, the independence of the self-regulatory body is not guaranteed under this model. Similarly, European institutions look set to adopt amendments to the Audio-visual Services Media Directive whereby national authorities would be tasked with determining – and sanctioning as appropriate - whether video-sharing platforms have adopted appropriate measures to combat 'hate speech' and protect minors from harmful content; see ARTICLE 19, New EU legislation must not throttle online flows of information and ideas, 12 September 2017.

[209] See, e.g., ARTICLE 19's response to the recognition of IMPRESS in the UK, 25 October 2017.

[210] In practice, there is no uniform definition of "self-regulation," however. Models labelled "self-regulation" in one country may qualify as "co-regulation" elsewhere.

[211] An example of "voluntary" codes adopted jointly with public authorities are e.g. the EU Code of Conduct on Countering Illegal Hate Speech; see, ARTICLE 19, EU: European Commission's Code of Conduct for Countering Illegal Hate Speech Online and the Framework Decision, August 2016; or the cooperation between the UK Counter-Terrorism Internet Referral Unit and the internet companies; see Metropolitan Police, 250,000th piece of online extremist/terrorist material to be removed, 23 December 2016.

[212] The Guiding Principles, *op.cit.*

[213] See, e.g., C. Mak, Fundamental Rights and Digital Content Contracts, Centre for the Study of European Contract Law Working Paper no. 2012-06, 2012.

[214] See, *mutatis mutandis,* European Court, *Dink v Turkey*, App. Nos. 2668/07, 6102/08, 30079/08, 7072/09, 7124/09, 14 September 2010. It is also consistent with consumer law principles; see, e.g., the recommendations of EU consumer protection bodies that Facebook, Twitter and Google have accepted to implement, 15 February 2018.

[215] For instance, in determining whether the contractual term in question was unfair or whether the social media platform or intermediary acted unconscionably e.g. in ending the contractual relationship based on its interpretation of general clauses of the contract regarding 'illegality', 'good morals' or 'public policy;' see C. Mak, *op. cit.*

[216] For more details, see also ARTICLE 19, Self-regulation and hate speech on social media platforms, 2018.

[217] This is also consistent with Manila Principles, *op.cit.*, Principle 1 (b).

[218] *C.f.,* the case involving the removal of a post containing the painting by Courbet *L'Origine du Monde*, in which French courts recently asserted their jurisdiction; see Facebook's nudity policy laid bare, 10 March 2015.

[219] *C.f.* Manila Principles, *op.cit.*, Principle 3 (c).

[220] *C.f.,* Manila Principles, *op.cit.*, Principle 3 (c).

[221] ARTICLE 19 will examine the privacy implications of Terms of Service in a separate policy.

[222] The Ranking Digital Rights indicators are available from here.