ARTICLE 19

Watching the watchmen

Content moderation, governance, and freedom of expression

2021

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.

E: info@article19.org
W: www.article19.org
Tw: @article19org
Fb: facebook.com/article19org

EXECUTIVE SUMMARY

In this policy, ARTICLE 19 outlines our position on social-media platforms' regulation of content moderation in a way that protects the right to freedom of expression and information.

The policy builds on our previous work in this area, in particular our policies on intermediary liability and on companies' community guidelines/terms of service. Previously, our proposals have been largely based on what might be understood as light regulation. We have argued that social-media platforms should continue to largely benefit from immunity from liability for the content of their users. This important rule does not prevent companies from being made accountable for failing to remove illegal content. At the same time, this model is predicated on companies operating terms of service in a way that is compatible with international standards on human rights. We have defended this model because we believe that it better protects the speech rights of users. It is important for not only social-media platforms but also a wide range of other Internet actors that deliver infrastructure-level services, such as content-delivery network services and domain-name registrars.

Today, however, this paradigm appears unsustainable in the face of the scale of the biggest social-media platforms and their consistent failure to appropriately address the criticisms that have been levelled against them – from Facebook's handling of the Rohingya crisis in Myanmar to YouTube's unfathomable position on 'hate speech' and the relentless attacks against women journalists on Twitter. We also believe that transparency should be a basic requirement that pervades everything companies do, accompanied by greater accountability and commitment to the protection of human rights.

In this policy, ARTICLE 19 examines whether the model we have been advocating for holds water in the face of the criticisms that have been made against the biggest social-media platforms. We examine the pros and cons of the various proposals that have recently been made around the world. Finally, we put forward our revised position on the regulation of platforms.

Key recommendations

1. States should refrain from unnecessary regulation of online content moderation.

2. Overarching principles of any regulatory framework must be transparency, accountability, and the protection of human rights.

3. Conditional immunity from liability for third-party content must be maintained, but its scope and notice and action procedures must be clarified.

4. General monitoring of content must continue to be prohibited.

5. Any regulatory framework must be strictly limited in scope. Regulation should focus on illegal rather than 'legal but harmful' content. Private-messaging services and news organisations should be out of scope. Measures should not have extraterritorial application.

6. Obligations under any regulatory scheme must be clearly defined. These include, in particular, transparency obligations and internal due-process obligations.

7. Any regulator must be independent in both law and practice.

8. Any regulatory framework must be proportionate.

9. Any regulatory framework must provide access to effective remedies.

10. Large platforms should be required to unbundle their hosting and content-curation functions and ensure they are interoperable with other services.

TABLE OF CONTENTS

INTRODUCTION

In 1996, John Perry Barlow, one of the fathers of the Internet, declared the Independence of Cyberspace.[1] His vision was one where 'anyone, anywhere may express his or her beliefs, no matter how singular, without fear of being coerced into silence or conformity'. Governments were not welcome there. Governance was to be derived from 'ethics, enlightened self-interest, and the commonweal'. Nearly 25 years later, the Internet is more commonly understood as comprising of a handful of social-media platforms (particularly Facebook, Google, and Twitter), censorship is rife, and governments are poised to regulate. What happened?

People's perception of the role of social-media platforms has also changed. In their early days, these platforms were widely seen as a powerful force for good, liberating free expression, enabling connections between people, and spearheading a democratic revolution across the world. Over the years, however, they have come to be viewed as a hotbed of 'hate speech', harassment, bullying, conspiracy theories, and propaganda. From merely 'hosting'[2] content in a relatively neutral way, they now actively promote selected third-party content, or even produce their own content. In the dock is the repeated failure of the biggest social-media platforms to grasp and address the concerns of their users and governments, from the Cambridge Analytica scandal to Facebook's failure to remove incitement to genocide against Rohingyas in Myanmar and YouTube's struggle to shut down the video of the Christchurch terrorist attack in New Zealand.

The amount of power these companies wield over individuals, and their dominance in several markets,[3] came into sharper focus in January 2021, when the major social media companies – including Twitter, Facebook, and many others – suspended the accounts of then-US President Donald Trump, due to the likelihood of future violence as a result of his tweets inciting storming of the US Capitol earlier in the month.[4] In the same month, Twitter locked the official account of the Chinese Embassy to the US, after a post that defended the Chinese government's policies in the western region of Xinjiang, where critics say China is engaged in the forced sterilisation of minority Uighur women.[5] Although not for the first time, this was a remarkable display of power on the part of social-media platforms. This has re-ignited one of the most fraught debates in content-moderation circles, i.e. who gets to decide what is or is not allowed in public discourse online.

This is an important debate to be had, and States have a role to play in it. In practice, much of the blame has been laid at the door of the regulatory frameworks governing the liability of Internet intermediaries. In many countries, social-media and other digital platforms have benefited from a regime of broad or conditional immunity from liability for hosting illegal content.[6] In the last few years, these regulatory frameworks have been under sustained attack as giving a free pass to these companies and helping them to grow profits on the back of algorithms that promote addictive engagement with 'extremist' and other 'harmful' content. This has raised the question of whether greater regulation is needed to tame the power of dominant social-media companies, tackle illegal and other harmful content, and provide greater democratic accountability for their decisions to the wider public. Most recently, governments have also responded with proposals that would seek to put social-media platforms under the purview of broadcast-type regulators, or proposed that the platforms should have a 'duty of care' to their users to prevent 'harm' caused by the speech of other users of the platform.[7]

For freedom of expression advocates, these proposals raise difficult questions about who we should trust with policing users' expression. While social-media platforms used to be perceived as providing a high level of protection to freedom of expression, they have

increasingly restricted their community standards, often silencing minority voices.[8] The transparency and dispute resolutions over content removals have so far been insufficient to enable sufficient scrutiny of their actions and provide meaningful redress for their users. Finally, it is doubtful that a small number of dominant platforms should be allowed to hold so much power over what people get to see without more direct public accountability.

At the same time, the prospect of new 'platform' regulation is deeply problematic. Many current proposals actually concern online 'content' regulation, i.e. the regulation of users' speech, as States effectively demand that companies police human communications and decide what speech is 'illegal' or 'harmful'.[9] This is deeply problematic as only the courts can determine illegality and different types of content may well call for different types of regulation; the solutions used to deal with child-abuse material may not be appropriate to deal with disinformation or copyright.

In any event, ARTICLE 19 believes that the new proposals for platform regulation require a careful examination from a freedom of expression perspective. To this end, States must identify the necessary and least restrictive method to achieve effective protection of each of the objectives traditionally assigned to media regulation (such as pluralism and diversity of freedom of expression), taking into account the evolution and roles of digital platforms in promoting and protecting human rights – including freedom of expression – online.

This policy offers this kind of analysis in light of international standards on freedom of expression. It builds on our existing position on intermediary liability,[10] our policy on internal rules/community guidelines of digital platforms and freedom of expression,[11] our proposals for a new model of multi-stakeholder regulation – Social Media Councils[12] – and our advocacy work towards the digital platforms. As we evaluate whether our long-held positions on these issues remain sustainable, we make recommendations as to what minimum safeguards a regulatory framework governing the activities of social-media platforms should include.

The scope of this policy primarily focuses on 'dominant' social-media platforms, which we sometimes refer to as 'digital platforms'. However, we are very mindful of the wider context of our inquiry, and we do refer to other players in the wider Internet ecosystem, including Internet infrastructure providers (such as computer delivery networks) and providers of other digital services (such as private-messaging or cloud services).

This policy is divided into four parts:

- First, it sets out some key terms, including 'dominant' social-media platforms, 'content moderation', and 'self-regulation;'
- Second, it outlines the applicable standards for the protection of freedom of expression online that should guide any legislative and policy efforts in this area;
- Third, it sets out the key arguments in favour of greater regulation over platforms' content-moderation policies and our overarching response to them;[13] and
- Finally, ARTICLE 19 makes recommendations as to what basic safeguards any new regulatory framework in this area should include to protect human rights.

This policy complements other ARTICLE 19 policies related to platform regulation, in particular our proposal on 'unbundling' content curation,[14] media diversity in the digital ecosystem,[15] and must-carry obligations.[16] We are also planning to issue further policy positions on other issues involved in this complex topic.[17]

KEY CONCEPTS

ARTICLE 19 has previously set out a typology of Internet intermediaries – internet service providers, web-hosting providers, social-media platforms, and search engines. 18 These remain relevant today, but as social-media platforms' practices have evolved, and their activities have come under greater scrutiny, it is important to define some key concepts in this debate. For the purposes of this policy:

- **Social-media platforms** are companies that enable individuals and groups to connect and interact with other users and to share content using electronic communication networks. In doing so, they provide 'hosting' services to users, but this is also generally accompanied by a range of other services, such as private messaging, 'content moderation', and the curation of news feeds. In addition, social-media platforms usually offer ad space and related services to advertisers and news-referral services to news publishers, connecting the latter with users; they can also provide application programming interfaces (APIs) to app developers, allowing them to program and develop tools that integrate with the platform. For this reason, some scholars have tentatively identified four relevant markets where social-media platforms usually operate: social-networking services, advertising display, news referral, and platforms for apps.[19]

- **Digital platform** is a term often used in this policy instead of 'social-media platforms'. However, digital platforms cover a wider set of activities than social media. They can include online marketplaces, apps stores, payment systems, search engines, and platforms for the collaborative economy. Digital platforms have been characterised, inter alia, as: (1) having a business model that is largely based on collecting, processing, and editing large amounts of data; (2) operating in multisided markets; (3) benefiting from network effects, where the value of the service tends to increase with the number of its users; and (4) relying on information-communication technologies to reach their users instantly at no cost.[20]

- **Content moderation** includes the different sets of measures and tools that social-media platforms use to deal with illegal content and enforce their community standards against user-generated content on their service. This generally involves flagging by users, trusted flaggers or 'filters', removal, labelling, down-ranking or demonetisation of content, or disabling certain features.

- **Content curation** is social-media platforms' use of automated systems to rank, promote, or demote content in newsfeeds, usually based on their users' profiles. Content can also be promoted on platforms in exchange for payment. Platforms can also curate content by using interstitials to warn users against sensitive content or applying certain labels to highlight, for instance, whether the content comes from a trusted source.

- **Self-regulation** is a framework that relies entirely on voluntary compliance; legislation plays no role in enforcing the relevant standards. Its *raison d'être* is holding its members accountable to the public, promoting knowledge within its membership, and developing and respecting ethical standards. Self-regulation models rely, first and foremost, on members' common understanding of the values and ethics that underpin their professional conduct. In the context of social-media companies, the term 'self-regulation' has been used to refer to 'solo-regulation' (the regulation of speech through terms of service or community standards) and sectoral regulation (the range of initiatives social-media platforms adopt to address particular issues, e.g. 'disinformation'). 'Self-regulation' is also sometimes used throughout this policy to describe government initiatives in which

companies are pressured to adopt a set of measures under government supervision, but where failure to comply with such measures cannot result in legal sanctions.[21]

- **Intermediary liability**[22] refers to the laws whereby Internet intermediaries can be held legally responsible for the content disseminated or created by their users. In many countries, Internet intermediaries are *immune* from liability (i.e. they cannot be taken to court) as long as they remove content once they obtain *actual* knowledge of illegality, or a court tells them to do so. Immunity from liability is generally considered essential to the survival of Internet intermediaries, given the scale of content that users produce.

- **Publishers' liability**[23] refers to the legal responsibility of anyone for the content they publish. In practice, this means that publishers are subject to laws of general application, e.g. defamation, laws preventing the publication of official secrets, etc. Newspapers are a prime example of this. Newspapers are legally responsible for the content produced on their website or in print. This includes letters to the editors. In practice, the content that newspapers publish, whether offline or online, is subject to thorough editing and legal vetting prior to publication. Nowadays, anyone can be a publisher thanks to the Internet. Internet users may therefore also be held liable for the content they produce.

APPLICABLE INTERNATIONAL HUMAN RIGHTS 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),[24] and given legal force through Article 19 of the International Covenant on Civil and Political Rights (ICCPR)[25] and in the regional treaties.[26]

The scope of the right to freedom of expression is broad and applies to all forms of electronic and Internet-based modes of expression. 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. Under international human rights standards, the legal framework regulating the mass media should take into account the differences between the print and broadcast media and the Internet,[27] while the telecommunications and broadcasting sectors could not simply be transferred to the Internet.[28] States should adopt a tailored approach to address illegal content online, and promote self-regulation as an effective tool in redressing harmful speech online.[29]

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 the respect of the rights or reputations of others, or the protection of national security or 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.[30]

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.[31]

Social-media companies and freedom of expression

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

Intermediary liability
The special mandates on freedom of expression have long maintained that immunity from liability is the most effective way of protecting freedom of expression online. 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 that is published online.[32]

Successive UN Special Rapporteurs on the promotion and protection of the right to freedom of opinion and expression (Special Rapporteurs on FoE) have stated several times that censorship should never be delegated to a private entity, and that States should not use or force intermediaries to undertake censorship on their behalf.[33] Further, David Kaye posited that 'smart regulation, not heavy-handed viewpoint-based regulation, should be the norm, focused on ensuring company transparency and remediation'.[34] Although he did not rule out the possibility of regulation under certain conditions, he reiterated that States should only seek to restrict content pursuant to an order by an independent and impartial judicial authority, and in accordance with due process and standards of legality, necessity, and legitimacy.[35] He also noted that States should refrain from imposing disproportionate sanctions – whether heavy fines or imprisonment – on Internet intermediaries, given their significant and chilling effect on freedom of expression.[36] His successor and the current Special Rapporteur on FoE, Irene Khan, has come to a similar conclusion.[37]

Human rights responsibilities of the private sector
Recommendations on social-media companies' responsibilities to respect human rights include (but are not limited to) the following instruments:

- **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.[38] They recognise the responsibility of business enterprises to respect human rights, independent of State obligations or the implementation of those obligations, and recommend several measures that companies should adopt. [39] These include incorporating human rights safeguards by design to mitigate adverse impacts, building leverage and acting collectively to strengthen their power vis-a-vis government authorities, and making remedies available where adverse human rights impacts are created.

- **Freedom of expression mandate holders** have addressed the role of social-media platforms in promoting freedom of expression and recommended that they respect and promote the Guiding Principles in this regard. They have recommended, *inter alia*, that companies should establish clear and unambiguous terms of service in line with international human rights norms and principles,[40] produce transparency reports,[41] and provide effective remedies for affected users in cases of violations.[42] They have also repeatedly called on States not to require the private sector to take steps that unnecessarily or disproportionately interfere with freedom of expression, whether through laws, policies, or extra-legal means, as they are typically ill-equipped to make determinations of content illegality.[43] The Special Rapporteur on FoE also made it clear that companies should embark on radically different approaches to transparency at all stages of their operations and should open themselves up to public accountability, suggesting that this could take the shape of Social Media Councils.[44]

- On the regional levels:

- The Council of Europe Commissioner for Human Rights recommended that States stop relying on Internet companies to impose restrictions that violate States' human rights obligations, [45] and that further guidance should be developed on the responsibilities of business enterprises in relation to their activities on the Internet.46 Similarly, the Committee of Ministers of the Council of Europe recommended that social-media companies respect human rights and the rule of law, including procedural safeguards,[47] and are transparent about their use of automated data-processing techniques.[48]

- The Organization of American States (OAS) has produced several reports on Internet freedom making similar recommendations.[49] For instance, the OAS has emphasised that intermediaries are still private entities, with interests that differ from those of the State, and that requiring them to function as a court goes beyond the scope of their competence and may provide incentives to abuse.[50]

- The African Commission on Human and People's Rights recently adopted the revised Declaration of Principles on Freedom of Expression and Access to Information in Africa.[51] Principle 39 of the Declaration provides that States must require Internet intermediaries to ensure that, in moderating and filtering online content, they mainstream human rights safeguards in their processes, adopt mitigation strategies to address all restrictions on freedom of expression and access to information online, ensure transparency on all requests for removal of content, incorporate appeals mechanisms, and offer effective remedies where rights violations occur.[52]

- **Civil society** has also made recommendations that social-media companies should respect international human rights standards. For instance, the Manila Principles on Intermediary Liability elaborate on the types of measures that companies should take to respect human rights. [53] In particular, companies' content-restriction practices must comply with the tests of necessity and proportionality under human rights law[54] and should provide users with complaints mechanisms to challenge companies' decisions.[55] The Ranking Digital Rights project assesses the major Internet companies for their compliance with digital rights indicators, which include *inter alia* availability of terms of service; reasons for content, account, or service restriction; notification to users; a process for responding to third-party requests; and various transparency activities.[56]

Content-specific principles

The human rights mandate holders have issued a number of joint declarations highlighting the responsibilities of States and companies in relation to specific content, including on 'violent extremism',[57] 'fake news',[58] online gender-based harassment and abuse,[59] and 'hate speech'.[60] In these, they have called on States not to subject digital platforms to mandatory orders to remove or otherwise restrict content, except where the content is lawfully restricted in accordance with international standards.[61] They have also called on companies to ensure that their users can easily access and understand their policies and practices, how they are enforced, and how they respect minimum due-process guidelines,[62] and to ensure that these policies are directly tied to international human rights law.[63]

The role of artificial intelligence in content moderation

In August 2018, in a report to the General Assembly, the Special Rapporteur on FoE made a number of recommendations regarding the role of artificial intelligence (AI) in content moderation and its potential impact on human rights. [64] In particular, he recommended that States should create a policy and legislative environment conducive to a diverse, pluralistic

information environment in the AI domain.[65] Such measures could include the regulation of technology monopolies to prevent the concentration of AI expertise and power in the hands of a few dominant companies, and regulation designed to increase interoperability of services and technologies.[66] For companies, he recommended, inter alia, that they should explicitly state where and how AI technologies are used on their platforms, services, and applications;[67] publish data on content removals, case studies, and education on commercial and political profiling;[68] and give individual users access to remedies for the adverse human rights impacts of AI systems.[69]

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.[70] The inability to communicate privately substantially affects individuals' freedom of expression rights.

This was recognised in several reports of David Kaye, the Special Rapporteur on FoE, in which he expressed concerns over States and private actors monitoring and collecting information about individuals' communications and activities on the Internet. These practices can constitute a violation of Internet users' right to privacy, and ultimately impede the free flow of information and ideas online.[71] The Special Rapporteur on FoE also recommended that States should ensure individuals can express themselves anonymously online and refrain from adopting real-name registration systems.[72] Further, he 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.[73] He also recommended that corporate actors reconsider their own policies that restrict encryption and anonymity (including through the use of pseudonyms).[74]

DILEMMAS OVER REGULATING CONTENT MODERATION

As noted earlier, in recent years, there has been increasing momentum for States to regulate the content-moderation activities of those digital platforms considered 'dominant' in the social-media market. In practice, these calls for 'platform' regulation often translate into 'online content' regulation, though the line is often blurred in public discourse.

In this section, ARTICLE 19 examines some of the core reasons why there is a perceived need for regulators to step in – and, in particular, to adopt an overarching framework for the regulation of online content. We then proceed to examine whether these reasons stand up to scrutiny.

Arguments in favour of platform regulation

Platform power and market power

By and large, the biggest factor driving the current impetus for regulation is the market power of a small number of social-media platforms, which leads to control over how a huge number of people exercise their right to freedom of expression online. These platforms' size and revenues are often bigger than those of several countries.[75] From the perspective of freedom of expression and democracy itself, the biggest concern is that only a handful of global – primarily US-based – companies decide what information users get to see, access, and share.

In practice, the concern over concentrated power has largely translated into a number of proposals that would place dominant social-media platforms under the purview of a broadcasting-type regulator. The focus of these proposals is therefore very much on online content regulation; they do not address market power directly.[76] At the same time, some elements of these proposals take into account the size of the biggest platforms, including by reference to their number of users and turnover.[77]

Other proposals involving the dissemination of content more explicitly aim to level the playing field between different types of industries.[78] In addition, competition and broadcasting regulators have started developing their own recommendations on how to deal with market failure in the provision of online services, including developing effective regulatory interventions and avoiding negative unintended consequences.[79]

Finally, it is worth noting that a number of data protection-related measures, from data portability to consent and purpose limitation, have a bearing on the competitiveness of platforms beyond the protection of users' personal data. The adoption of strong data-protection frameworks, such as the General Data Protection Regulation, is therefore key to keeping the power of these mammoth companies in check and putting limits on their extractive business models.

Although most proposals for a new regulatory framework more squarely aim to tame the excessive power of the largest US social-media platforms, they have also been used as an opportunity to regulate smaller players and other services, such as messaging, cloud-hosting, and file-sharing services. Indeed, some infrastructure providers have increasingly been drawn into the debate about content moderation, in particular over providing their services to neo-Nazi and far-right groups.[80]

Platform business model

The second key concern underlying the current drive for greater online content regulation is the business model of the biggest social-media platforms, based on the collection of vast amounts of data about their users and their online habits (behavioural data) and its monetisation through online (targeted) advertising.[81] Privacy and digital rights advocates, in particular, have long warned how 'surveillance capitalism' is both manipulative and harmful to users' human rights.[82] It relies on techniques that are both highly privacy-invasive (e.g. online tracking, real-time bidding, and targeting) and incredibly opaque.[83] These techniques seek to keep users engaged on platforms and fuel the widespread dissemination of clickbait, sensationalist, and 'extremist' content online, without users having a meaningful understanding of why they are seeing a particular type of content or ad on social media.[84] These advertising techniques, therefore, impinge on users' informational self-determination. They also tend to promote lower diversity and quality of content,[85] and present significant risks of discrimination in areas such as housing and job advertising, among others.[86]

Perceived failure of solo-regulation and self-regulation, and lack of democratic accountability

Another key driver of current regulatory proposals is the perception that the major social-media platforms have failed to address a number of 'harms' – from 'disinformation' and 'hate speech'[87] to terrorism and child-abuse images. In particular, critics point to the widespread availability of both illegal or 'harmful' material online – including its sometimes-tragic impact offline[88] – and the lack of consistency and transparency in social-media platforms' approach to content removal.

Much of the blame has been pinned on the prevailing model of conditional-immunity framework, which is perceived to have given platforms a get-out-of-jail-free card. They are seen as profiting off the back of social ills, which they have enabled, without any public accountability. As such, the various initiatives that large social-media platforms have adopted of their own volition.[89] or in response to 'self-regulatory' codes of practice.[90] are regarded as insufficient to both tackle these problems and bridge the accountability gap.[91]

Need for legal certainty and real transparency

Faced with the challenges outlined above, and a myriad of different laws to deal with them, some governments have argued that an overarching regulatory framework would bring more clarity and legal certainty.[92]

Moreover, these frameworks are considered necessary to force social-media companies to be more transparent about their content-moderation operations and how their algorithms work. Beyond statistics, it remains highly unclear what content gets removed, for what reasons, and why users get to see particular types of content. Under these new frameworks, companies would also be required to put in place redress mechanisms for wrongful removals of content. These new obligations would be overseen by a regulator that would not otherwise be involved in decision-making about content.[93]

Lack of media diversity

Finally, another potential reason for regulators to step in is the protection of media pluralism and diversity in the digital ecosystem. Some legislators have already put forward proposals to remedy what they perceive to be social-media platforms' failure in this area.[94] This is a concern for three main reasons:

- **Excessive concentration in the social-media market and bottlenecks**: A very limited number of large platforms gets to decide how news is distributed on their service and the criteria that apply to such distribution. The largest platforms are able to do so because they concentrate the greatest numbers of users. Media companies therefore have no choice but to engage with them, largely on their terms, to get users' attention.[95]

- **Personalised content**: The very model of social-media platforms implies that their users are only exposed to highly personalised content, rather than a diversity of viewpoints.[96] This has led to fears of filter bubbles[97] and accusations of bias from some political parties.[98]

- **Domination of online advertising**: Through online advertising, traditional media has suffered not only losses of audiences to 'free' online sources and social media but also losses in advertising revenue. The problem of media sustainability also poses a threat to media pluralism and diversity.

ARTICLE 19's position on arguments for platform regulation

ARTICLE 19 recognises that the concerns behind the drive for greater regulation of platforms' practices are entirely valid. Indeed, we share them. However, we believe that many of the solutions currently being proposed are likely to miss the mark, entrench the 'dominance' of the largest players. and be open to abuse by governments by giving them more control over platforms and content. Instead, we outline below some solutions that, we believe, would better guarantee the protection of freedom of expression.

The concentration of power in the hands of a few large platforms should be addressed primarily by pro-competition tools

In ARTICLE 19's view, the most recent developments in 'platform regulation' point to the complexity of the challenges thrown up by large digital platforms and the extent to which various disciplines – from online content regulation to competition, consumer protection, and data protection – are interrelated. This also demonstrates that any regulatory intervention must be clear about its objectives and the 'harms' it seeks to address. In particular, ARTICLE 19 believes that concerns around the size, power, or 'dominance' of social-media companies ought to be addressed primarily by competition tools and pro-competitive economic frameworks, rather than online content regulation.

At the same time, we recognise that some social-media companies have become so dominant that it may be appropriate to adopt tiered approaches. Whereas the largest players will generally be able to meet new regulatory obligations (e.g. on transparency, due process, or even content-removal targets), small players would simply go out of business trying to meet them. To put it differently, it is vital that smaller companies are not unduly burdened by heavy regulatory obligations; otherwise, they would be unable to compete with the biggest companies. We make recommendations on how this ought to be addressed later in this policy.

We note, however, that tiered approaches should be designed in a way that avoids the undesired effect of entrenching the position of incumbent actors, as they would be able to cope with new regulatory requirements that could make their services more attractive to users (e.g. safety and trustworthiness).[99] We also note that tiered approaches might have the unintended effect of driving 'bad actors' to less-policed platforms;[100] as such, 'bad actors'

would still be operating on the margins, with less visibility to the public but more visibility among themselves.

The effectiveness of stricter regulatory approaches has not been established

ARTICLE 19 also believes it is not clear whether new State regulations on platforms' content moderation are strictly necessary to tackle problematic content, such as 'hate speech' or terrorism, online.

We are concerned that new legislation often fails to comply with international human rights law.[101] There is also often a lack of evidence of proposed measures' effectiveness, whether in terms of combating specific issues (such as 'extremism' or 'hate speech')[102] or their collateral damage on freedom of expression.[103] All too often, lawmakers seek to adopt laws to send a political message to the public that 'something is being done' to address an issue, rather than investing resources in less visible but more effective long-term solutions.

For all these reasons, ARTICLE 19 remains deeply sceptical that the types of new laws that have been put forward can offer solutions to counter problematic content; they are more likely to be counterproductive, aim at the wrong target, and be detrimental to freedom of expression.[104]

Independent or multi-stakeholder governance models should be set up instead

ARTICLE 19 also recognises that social-media companies have generally been slow to react to a number of legitimate concerns – such as concerns over 'hate speech', 'disinformation', and the protection of children – on their platforms. Over time, this has considerably eroded public trust in them and raised legitimate questions about their lack of public accountability.

Nonetheless, in our view, it is too early to dismiss the measures and initiatives they have adopted to tackle these problems as ineffective. Although they still have a long way to go, social-media companies have responded by adopting a range of measures, which have become more sophisticated over time. These include regularly updating and clarifying their policies,[105] investing in the upgrade of their systems to help them detect problematic content in a wider range of languages,[106] expanding their fact-checking[107] or trusted-flagger programmes to a wider range of countries to better understand national specificities,[108] and developing more sophisticated responses to dealing with 'problematic' content. They also consistently seek to innovate and regularly develop new features giving users more control over what they see,[109] and have put in place mechanisms to appeal against wrongful removals of content based on their community standards.[110] Some companies have started to undertake human rights assessments,[111] and continue to increase the number of human moderators on their platforms,[112] often in response to particular crises or anticipated regulatory moves.

Moreover, many areas of problematic content that drive the push for platform regulation are complex and hard to define. For example, 'disinformation' is a complex problem – not least because, more often than not, the line between fact and opinion is hard to draw. For this reason, banning and other legal restrictions on sharing false information are open to abuse and can have a devastating impact on political discourse. ARTICLE 19 notes that the impact of social-media companies' policies in this area remains largely unknown.[113] Overall, we believe there is little evidence that current initiatives have either succeeded in or failed to address concerns around spreading 'disinformation'. Moreover, some academic studies have pointed out that the extent to which disinformation is indeed 'harmful' to democracy is highly speculative.[114] For all these reasons, we believe that the case for greater regulation in this area

is not borne out.[115] At the same time, we recognise that there could be a case for independent public oversight or auditing of the claims social-media platforms make about how they tackle 'disinformation'.[116]

Further, many types of problematic content are difficult and complex issues with no easy solutions. For instance, a number of problems that 'hate speech' poses have undoubtedly been exacerbated online, as this type of content is prone to go viral. It is both expected and necessary that governments and companies should take action to tackle these problems. The question is whether the measures that social-media companies have adopted so far are sufficient – and, if not, whether new legislation tackling 'hate speech' online is an appropriate solution to the problem at hand. Similarly, the steps social-media companies have taken to tackle 'hate speech' should not be discarded out of hand.[117]

Hence, it is difficult to conclude that self-regulatory initiatives have entirely failed. This does not mean that social-media companies can afford to become complacent: they must continue to do more to tackle 'hate speech' and be more transparent about it. It is particularly important that the various stakeholders in this debate, including governments, companies, and civil society, continue to dialogue to ensure the right balance between the protection of the rights to equality and freedom of expression.

In this respect, ARTICLE 19 believes that new forms of multi-stakeholder or independent self-regulation models, such as Social Media Councils, could be part of the solution.[118] This is especially true in some areas, particularly where the content at issue is not illegal. Hence, some forms of independent self-regulation remain necessary for the following key reasons:

Arguments for an overarching regulatory framework are problematic

ARTICLE 19 notes that the other arguments put forward in favour of an overarching online-content regulatory framework are highly debatable. In particular:

- **The need for legal certainty:** This is a powerful argument in support of a more unified regulatory framework. This is the case, for instance, in the EU.[119] At the same time, ARTICLE 19 notes that, as far as the EU is concerned, it seems hard to reconcile with its approach so far, i.e. the adoption of separate legal instruments to deal with copyright or terrorism online, and the revision of the Audio-Visual Media Services Directive to deal with 'hate speech' and content harmful to children online.[120] In many instances, the measures the EU has adopted in the last few years have not been implemented yet. It is therefore unclear whether they are effective. Overall, we believe that the concomitant adoption of a Digital Services Act may not necessarily provide greater clarity in this area of law. It is also highly questionable whether a single regulator, whether at domestic or European level, would be well placed to deal with concerns as diverse as 'disinformation', 'hate speech', child-abuse images, and terrorist content.

- **Regulation of a wide array of services:** Furthermore, as we explain in more detail below, we are concerned that reform in this area could be used to regulate services or actors such as private-messaging apps or the press, which should be out of scope due to the threat such regulation would represent to the protection of the rights to freedom of expression and privacy.

- **Setting up a one-stop-shop regulatory body:** Finally, although we recognise that a regulator might contribute to the enforcement of transparency and due-process obligations, we are concerned that process questions are often entangled with content-related ones, such as the adoption of filters to detect harmful or illegal content.

> Regulators also remain a powerful avenue for governments to exercise control over information flows, particularly in countries where the regulator is not independent. Giving regulators powers over the provision of social-media services could ultimately have a chilling effect on users' freedom of expression.

Hence, ARTICLE 19 is not convinced that the case for a regulator overseeing online content and the removal of immunity from liability has been made, particularly for content that falls into the category of 'harmful but legal'. We believe that other mechanisms, such as independent multi-stakeholder bodies (Social Media Councils), could be better suited to oversee social-media companies' compliance with a set of principles derived from human rights standards.

The lack of media diversity needs to be addressed

As a freedom of expression organisation, ARTICLE 19 takes concerns about media pluralism and diversity very seriously. While concerns over filter bubbles may have been overplayed,[121] it is equally clear that algorithms used for content curation are not neutral to diversity, and that changes to them are highly likely to have an impact on news consumption. There is also some evidence to show that *online* audiences are more polarised, which might, in turn, incentivise the production of more partisan content.[122] Ultimately, the overarching concern remains that certain platforms or search services hold too much power over the terms of the distribution of news – and, ultimately, the survival of journalism – due to their size, market share, and user base. ARTICLE 19 believes that some of these concerns could be addressed by a number of measures, including Social Media Councils (which would foster dialogue between platforms and news organisations) and the unbundling of hosting and content-curation services (which would bring greater competition, and therefore diversity and user choice, to content curation). We outline what these proposals might look like in more detail below.

The platforms' business model needs reform to comply with data-protection legislation

ARTICLE 19 believes that the business model of large social-media platforms and other companies significantly interferes with the right to privacy and can have a chilling effect on freedom of expression. At a minimum, it is clear that social-media companies must comply with data-protection legislation to restrict the amount of personal data they collect about users. ARTICLE 19 also notes that, in light of recent developments (particularly around elections), current political-advertising practices must be reviewed for their impact on freedom of expression, and some regulation might be needed in this area. ARTICLE 19 addresses this issue in a separate policy.[123]

KEY CONCEPTS AND FEATURES OF CONTENT-MODERATION REGULATION

In the space of a few short years, the broad consensus in Western countries that conditional immunity from liability was a necessary prerequisite to technical innovation and Internet freedom has been increasingly questioned. Likewise, the prohibition on general monitoring of content on social media platforms has been steadily undermined. As noted earlier, a number of States have gradually adopted or proposed a raft of measures that erode these important foundations of the protection of freedom of expression online.[124] In this section, we review some of the key concepts put forward in legislative proposals in this area.

At the outset, ARTICLE 19 notes a number of recurring misconceptions in debates about online content regulation:

- **Liability vs regulation:** Liability rules enable individuals to sue natural or legal persons for causing them harm on the basis of existing causes of action. Conversely, immunity from liability in the digital sector means that digital companies are protected from lawsuits unless they fail to take action upon obtaining actual knowledge of illegality. Given the millions of pieces of content posted every minute, and the potential for this content to be illegal, immunity from liability is a very important protection for companies – and freedom of expression more generally. If this protection fell away, companies would probably cease to exist altogether, given the legal risk they would run in allowing content to be posted on their services in the first place. This would also be detrimental to freedom of expression. Regulation, by contrast, means that companies have to comply with a defined set of obligations laid down in law and, if they fail to comply with those obligations, they may face sanctions such as fines. It is possible for both immunity from liability and regulation to coexist.

- **Content vs platform regulation:** The current debate around how digital companies are regulated often involves references to 'content regulation' or 'platform regulation'. It is important to remember that content is primarily regulated by law that applies directly to individuals, e.g. defamation law or the criminalisation of incitement to violence. The situation of companies is different, since they allow third parties (i.e. users) to publish content without prior vetting (though this is increasingly changing with the use of filtering technology). While current legislative proposals are presented as regulating *platforms*, they are largely delegating the responsibility of regulating *users'* speech to platforms, often beyond the requirements of the law (see further below). Moreover, it is important to bear in mind that, in practice, regulation is aimed at the *activities*, *behaviour*, and *services* of companies, rather than at the companies per se.

- **Publisher's liability vs intermediary liability:** Digital companies are increasingly labelled as 'media companies'. Some argue that the same liability that applies to publishers should also apply to dominant social-media platforms such as Facebook, Twitter, and YouTube. Although it might be said that some social-media platforms are increasingly behaving like traditional incumbent media companies, some key differences remain. In particular, they have millions of *users* posting content on their platform. Users are the primary publishers of this content. Social-media platforms select and organise that content. As such, they exercise a form of editorial responsibility – a process sometimes referred to as 'content curation'. However, they are not responsible for the content itself unless they modify it sufficiently that it might be said to be their own. The key question is whether the mere use of algorithms and filters is sufficient for tech companies to obtain

actual knowledge of illegality on their platforms, and therefore lose immunity from liability. We answer this question in the next section.

Positive features of legislative proposals on content-moderation regulation

ARTICLE 19 notes that several regulatory proposals in this area include some positive aspects in terms of protection of the right to freedom of expression, including:

- **Enhanced transparency requirements** about how social-media and other companies are handling content moderation, including transparency over the removal of users' content removal or transparency over handling complaints from their users.[125]

- **Requirements to put in place redress mechanisms and various other procedural safeguards,** such as notification of content takedowns to users whose content has been removed, and users being able to challenge companies' decisions before the courts.[126]

- **Immunity from liability:** Most proposals[127] do not explicitly repudiate immunity from liability. However, they put it at risk by mandating (more or less explicitly) 'proactive measures' to detect illegal or harmful content. The question becomes: Do companies retain enough of an incentive to protect freedom of expression if they are guaranteed immunity from liability through a 'Good Samaritan clause',[128] including when they put in place filters and other measures? On balance, for reasons we explain further below, we believe the answer to this question is 'yes'.

- **Systemic approaches:** While individuals retain the ability to sue companies for failing to take down 'illegal' content, current regulatory proposals do not propose to sanction single failures to remove content; rather, they take a more 'macro' approach to content moderation and focus on companies' systems and processes.[129]

- **Tiered approaches:** To the extent that such regulation is thought necessary – which we do not – a positive feature in many proposals is that they tend to provide for tiered approaches depending on the size of the company at issue, at least by reference to the number of users.[130]

Shortfalls of regulatory proposals on content moderation

Notwithstanding several positive features, most of the current proposals on content-moderation regulation create significant problems from the perspective of international legal standards on the protection of freedom of expression. As highlighted earlier, for an interference with the right to freedom of expression to be justified under international law, it must be provided by law; pursue a legitimate aim, as exhaustively listed in human rights treaties; and be necessary and proportionate to that aim. The vast majority of proposals fail to meet both the legality and proportionality tests.

Focus on 'legal but harmful' content

Most pending proposals go beyond the scope of the subject matter as they cover content which is not just *illegal* but also merely *harmful*.[131] This is deeply concerning because it means that companies are likely to be legally *required* to take measures against content that is

effectively allowed under the law. Individuals would be allowed to say things on the street that are not permitted on social-media platforms. Moreover, the vast majority of 'harmful' content is inherently hard to define, and is therefore likely to lead to the censorship of content that – though offensive to some – ought to stay up. This is the case, for example, for material having the 'likely' effect of humiliating a person,[132] or 'extremism' material 'with a less clear definition'.[133] Even if some of the proposals provide greater legal certainty (since they concern illegal material), the laws to which they refer raise concerns from the perspective of freedom of expression.[134]

Overbroad scope in the range of companies covered

Another significant concern is the extent to which a wide range of services, such as private-messaging or news sites, are likely to fall within the scope of existing proposals[135] when they really ought to remain out of scope. ARTICLE 19 is especially concerned that regulation in this area may require companies in scope to filter content, which would significantly weaken encryption and therefore the privacy of users' communications. For news sites, our concern is that they may become subject to regulation that could be used to silence their reporting, particularly when it is critical of politicians. Moreover, in existing proposals, little thought seems to have been given to the extent to which different service providers may be affected in different ways; for instance, an infrastructure provider (e.g. Cloudflare) ought not be made subject to the same obligations as a social-media platform (e.g. Facebook).

While some proposals tend to take a tiered approach to the obligations imposed on companies in scope, whether by reference to the number of their users or turnover, this is not always the case.[136] Moreover, there is little consistency in the criteria being used; exceptions are not always provided for non-profits (e.g. Wikipedia), and smaller providers with a large user base but low turnover may still be made subject to onerous obligations.[137]

Vague and problematic obligations

While some proposed obligations are generally positive (e.g. around transparency and due process), others are either too vague or deeply inimical to the rights to freedom of expression and privacy:

- **Obligation of duty of care:** Some proposals refer to a general 'duty of care' without clearly defining the types of obligations it entails[138] or its relationship with intermediary liability. In general, the 'duty of care' approach is based on an assessment of a risk of 'harm' to users. Even if a 'duty of care' entailed an assessment of the risks of 'harm' arising from companies' systems,[139] the term 'harm' is difficult to define, and defining what constitutes 'unacceptable' risks would fall within the regulator's discretion.[140] Without more clarity, a 'duty of care' could look like any of the obligations detailed further below, i.e. content removal within unduly short timeframes, proactive measures, or requirements to remove encryption.[141]

  ARTICLE 19 observes that proponents of a mandatory 'duty of care' seem to suggest that their proposals align with the type of human rights impact assessments encouraged by the Guiding Principles.[142] In this analysis, the suggestion appears to be that risk assessments concerning possible human rights impacts are the same as risks assessments of potential 'harm'. ARTICLE 19 believes that this is fundamentally mistaken. Human rights are legal concepts, contained in national laws, constitutions, and international instruments. Although they are typically cast in broad terms, they are further defined by a vast body of decisions by courts and tribunals. Consequently, despite the appearance that they are very broad, they are actually carefully defined. The concept

of 'harm', by contrast, has no general legal definition. To the extent that it *is* used in legal settings, it tends to be carefully defined and very context-specific, e.g. the criminal offence of occasioning actual bodily harm. The Guiding Principles are designed to prevent violations of human rights, not to prevent 'harm'. Moreover, contrary to the 'duty to prevent harm', the Guiding Principles do not require companies to prevent human rights violations *by others* as such.[143] At most, they require companies to adopt procedures that aim to avoid or minimising *their* involvement in human rights violations.

- **Obligation to remove certain content:** The vast majority of proposals effectively delegate censorship powers to private companies to remove illegal or harmful content. If they fail to do so, the regulator can order these services to be blocked.[144] Although some proposals include a requirement for the oversight body to ensure that companies do not *excessively* remove content,[145] there is generally no corresponding transparency requirement about *wrongfully* removed content. In any event, unless a complaint is made, the number of wrongfully removed pieces of content may never be known. Moreover, it is partially dependent on the effectiveness of both internal complaints mechanisms and other forms of judicial redress. However, most proposals are entirely silent on how the latter should be funded.

- **Obligation to remove content in short timeframes:** In general, most proposals tend to require the removal of 'hate speech' or 'manifestly unlawful' content within 24 hours and terrorist content with one hour.[146] These timeframes are clearly too short to make a proper – let alone legal – assessment of the claims being made.

- **Obligation to take proactive measures:** Some proposals contain the requirement for companies to adopt 'proactive' measures or make 'best efforts' to address illegal or harmful content. Sometimes this obligation is presented as 'recommendations' the regulator has made, though the law makes it clear that compliance with these recommendations is expected.[147] In practice, 'proactive measures' and related terms mean the adoption of filters or other technology to identify or prevent the upload of 'problematic' content. Filters cannot, of course, assess the legality of content, and have consistently been shown to be prone to error – particularly when an analysis of the context is necessary, such as in the case of 'terrorist' or 'hate speech' content. As a result, legitimate content may be wrongfully removed.

The regulator

Most proposals to regulate social-media platforms would place them under the purview of a 'broadcasting' regulator, which would see that regulator's remit and powers considerably expanded. For ARTICLE 19, this raises several concerns:

- **Necessity and legitimacy:** At the outset, we note that a primary concern with these proposals is that they would ultimately put users' speech under the control of a public authority that could be granted vast discretionary powers as to what amounts to 'harmful' content. Whether such a move is necessary, and its likely impact on Internet users' right to freedom of expression, remains unclear. In any event, it is widely inappropriate for a private company or regulator to determine the legality of content; this is a matter for the courts.

- **Independence:** A further key concern for a global organisation such as ARTICLE 19 is that the regulator may not be independent, whether in law or practice. This is not a theoretical concern, even in established democracies,[148] but also in many countries around the world where the rule of law is weak or under threat.

- **One-stop shop:** Finally, we are concerned that current proposals would place a vast array of different types of content under the purview of a regulator that may not be best equipped to deal with these issues, let alone provide the sort of solutions that would help resolve these problems. For instance, the removal of child-abuse material is a specialised area that may involve the need for special procedures and considerations, which would be best served by a dedicated institution. In our view, a 'one-stop shop' regulator for all illegal or 'harmful' content is unlikely to be effective.

Disproportionate sanctions

ARTICLE 19 notes that the range of sanctions currently considered in various countries, such as France or the UK, is particularly high, e.g. up to 4% of companies' global turnover. Coupled with unduly vague obligations, we worry that this is likely to give them an incentive to remove more content to the detriment of the protection of freedom of expression online. Moreover, we note that the threshold at which sanctions become likely is often unclear[149] and these measures are disproportionate.

Other concerns

ARTICLE 19 is concerned that platform-regulation laws or draft laws increasingly contain requirements for companies to appoint points of contacts or representatives 'in country'. In practice, this enables the authorities to put significant pressure on companies to comply with their demands, even if those have little to no legal basis and plainly seek to silence governments' critics.[150]

We also query the extent to which companies with global operations may be able to comply with different – and potentially contradictory – regimes across the globe. In our view, this could lead to a more fragmented Internet, and potentially a race to the bottom, with controversial content increasingly being removed by default under companies' terms of service.

Finally, we note that laws in this area are frequently adopted without a proper participatory process in which all sectors, including small and local digital platforms, are included.

ARTICLE 19'S RECOMMENDATIONS

Recommendation 1: States should refrain from unnecessary regulation of online content moderation

ARTICLE 19 makes clear that we do not support the models of regulation that we have seen emerging in Western Europe or Latin America. We remain concerned that the proposed rules are often overly vague and premised on the independence of a bona fide regulator. This simply is not the case in many parts of the world. The broader context in which regulatory proposals are made is particularly important.[151] Moreover, under the three-part test for restrictions on freedom of expression, States should comprehensively demonstrate that there is sufficient evidence to conclude that regulation is necessary. While the urge to regulate very large platforms is understandable, these platforms will be able to adapt and meet regulatory requirements. This is unlikely to be the case for small or medium-sized platforms. If nothing else, more online regulation is likely to benefit very large companies and entrench their dominant position.

We therefore urge lawmakers to resist the temptation of unnecessary regulation. If our societies' biggest concerns stem from the power and dominance of a small number of platforms, in our view, lawmakers should address those concerns using the remedies available under competition law and pro-competitive economic regulation. It is also for this reason that we believe our positions and recommendations on intermediary liability and content-moderation rules of social-media platforms still stand.[152]

In addition, ARTICLE 19 has advocated for oversight of social-media companies by an independent multi-stakeholder institution, such as Social Media Councils. We continue to believe that the bottom-up creation of an institution would lead to better outcomes for freedom of expression by precluding overly harsh sanctions. It would also foster more effective accountability, since its stakeholders would take ownership of the organisation's success. Social Media Councils and regulation of very large platforms through competition law can have complementary roles.

Recommendation 2: Overarching principles of any regulatory framework must be transparency, accountability, and the protection of human rights

Notwithstanding the above, ARTICLE 19 recognises that greater regulation of the digital sector, and social-media companies in particular, appears all but inevitable.[153] For this reason, we believe that the objectives of any new regulatory framework cannot be limited to only 'tackling illegality'. Equally, the 'prevention of harm' is much too broad a concept to be a meaningful, let alone a legitimate objective of any such framework.

Rather, we believe that the overarching principles of regulation must be transparency, accountability, and the protection of human rights. The latter means that the legality and proportionality principles must be upheld throughout. In addition, any such framework must be based on robust evidence to adopt the most appropriate solutions.

Recommendation 3: Conditional immunity from liability for third-party content must be maintained – but its scope, and its notice and action procedures, must be clarified

ARTICLE 19 notes that removing – or unduly limiting – immunity from liability would give digital companies an incentive to either filter and remove as much of users' content as possible or to be entirely neutral and not remove any content at all. In other words, it would either lead to increased censorship or remove any incentives for companies to engage in content moderation. In our view, both these outcomes would be highly undesirable for the protection of freedom of expression online.

For this reason, we believe that digital and social-media companies must continue to benefit from broad, or at least conditional, immunity from liability. In this respect, we note that, by definition, liability can only ever attach to illegal – rather than 'harmful' – content. In practice, this means that companies should not be held liable for failing to take undefined 'reasonable steps' to address 'harmful' content.

The liability must differ for different types of activities and services as follows:

Broad immunity from liability for those providing essential infrastructure services, including 'mere conduit' and neutral 'hosting', should be maintained

ARTICLE 19 believes that companies providing essential infrastructure services for the functioning of the Internet, such as content-delivery networks, should benefit from broader immunity from liability than services engaged in content moderation at the application layer. They should only be required to remove content by order of a court. In practice, this means that infrastructure providers should not be penalised for hosting certain websites, unless they have failed to comply with a valid court order requiring them to discontinue their services to such a website because it is illegal.[154] Equally, they should not be required to host such a website if they do not wish to do so. In other words, essential infrastructure providers should not be mandated to carry content.

At the same time, services providing an essential service should clearly set out the reasons why they may decide to discontinue the provision of services to certain actors.

Notice and action procedures for those providing hosting services coupled with content moderation

ARTICLE 19 further believes that the current standard of knowledge required to benefit from immunity from liability must be maintained, i.e. it should remain 'actual' rather than 'constructive' knowledge. ARTICLE 19 continues to believe that actual knowledge of *illegality* can only be obtained by a court order. To hold otherwise would be to accept that content is illegal simply because a third party, such as a copyright holder, said so.

At the same time, we recognise that a regulatory framework could clarify the different types of notice and action procedures applicable to different types of content – without prejudice to the determination of legality. In other words, the law should clarify how companies obtain 'actual knowledge' of alleged illegality, and what they ought to do about it once they obtain it.

ARTICLE 19 has previously set out how this could work in practice.[155] We believe that our suggested processes remain valid and provide the best way forward to protect the right to freedom of expression. In summary, this includes:

'Notice to notice' for private disputes, such as copyright or defamation

Under this procedure, the complainant or 'trusted flagger' would be required to give their name and set out in a notice why they believe their rights have been infringed, the legal basis for their claim, the location of the allegedly infringing material, and the time and date of the alleged infringement. The hosting provider would be required to pass on the notice to the alleged wrongdoer (i.e. the content provider) as soon as practicable but within a maximum period of time (e.g. 72 hours). The content provider would have a choice to remove the content or file a counter-notice within a reasonable period of time (e.g. 14 days).

Again, the hosting provider would be required to pass on the counter-notice as soon as practicable but within a maximum period of time (e.g. 72 hours). The complainant would then be given a period of time (e.g. 14 days) to decide whether they want to take the matter to court. The content would be removed following a court order. A hosting provider could be held liable for statutory damages if they failed to comply with their 'notice to notice' obligations, or if they failed to remove the content following a court order. By contrast, if the content provider failed to respond or provide a counter-notice within a given period of time, the hosting provider would lose immunity from liability. They could either remove the allegedly unlawful content or may be held liable for the content at issue if the complainant decides to take the matter to court or another independent adjudicatory body. To protect freedom of expression, any new 'notice to notice' framework should also provide for penalties for abusive notices.

'Notice and action' for allegations of serious criminality

Under this procedure, a hosting provider would be *required* to take down content when it receives a court order to that effect. In other words, they would be liable for failing to comply with such an order. In practice, this would mean that, if law-enforcement authorities believe that a piece of content should be removed and the matter is not urgent, they should seek a court order, if necessary on an ex parte basis. If, however, the situation is urgent (e.g. someone's life is at risk), law enforcement should be given statutory powers to order the immediate removal or blocking of access to the content at issue. However, any such order should be confirmed by a court within a specified period of time (e.g. 48 hours). The use of informal mechanisms – e.g. phone calls or emails requesting the host to remove content – should not be permitted.

By contrast, if hosting providers receive notice from an ordinary user about suspected criminal content, the host or platform should, in turn, notify law-enforcement agencies if they have reason to believe the complaint is well-founded and merits further investigation. The host or platform *may* also decide to remove the content at issue, as an interim measure, in line with their terms of service. However, they would not be required to do so, and failing to remove the content at issue would not attract liability.

The same process would apply to private bodies that work with law-enforcement agencies and operate hotlines that individual Internet users can call if they suspect criminal content has been posted online.[156] In other words, the hotline would report the content at issue to both the host and law-enforcement agencies. The host would use the same process it uses for complaints from ordinary users, i.e. it would remain free to decide whether to remove content on the basis of its terms of service. The same model could be applied to other bodies, whether public or private, that receive complaints from the public concerning potentially criminal content online, or to notice issues by 'trusted flaggers' (see below for further details on trusted-flagger programmes). Whichever option is pursued, it is important that the authorities are notified of any allegation of serious criminal conduct so that it may be properly investigated and dealt with according to the established procedure of the criminal justice system.

The Manila Principles on Intermediary Liability provide further useful guidance on how 'notice and action' procedures should work.[157] We believe that this is the most proportionate and rights-respecting way in which 'notice and action' procedures can be operated, particularly against small companies.

Protection from liability in case of content-moderation measures applied by companies of their own motion

ARTICLE 19 believes that social-media platforms and other digital companies should not be held liable simply because they have adopted community standards and use human moderators or other tools to enforce them.[158] In this sense, we support the adoption of a 'Good Samaritan clause' that would encourage content-moderation efforts made in good faith. In our view, failure to do so would prevent the adoption of innovative technical solutions and tools, such as demonetisation or the removal of certain platform features, that would strike a more proportionate balance between the protection of freedom of expression and tackling illegal – or even 'harmful' – content. At the same time, companies that use these tools should be subject to stringent transparency and due-process requirements about how they use them.

Similarly, companies should benefit from broad immunity from liability for the recommendations their algorithms make, in circumstances where those algorithms recommend illegal content in response to content users have viewed. While system developers and coders define the parameters within which algorithms operate, they do not control or determine the outcome of these automated processes. Algorithms produce results from datasets in ways that are both complex and unpredictable. They are also both generally prone to making mistakes and unable to distinguish between lawful and unlawful content. Holding companies liable for every possible 'mistake' their systems make would therefore be both unworkable and disproportionate. Insofar as liability deals with specific instances of illegality, it is also a poor instrument to address the systemic challenges thrown up by algorithms.

Instead, companies – particularly those with significant market power – should be subject to greater transparency obligations and required to carry out human rights impact assessments, as outlined below. In our view, the same reasoning should apply to navigation or 'findability' services, i.e. they should not be penalised if their search-engine algorithm returns illegal content, but they should be transparent and explain to the public how their algorithm functions to return search results.

By contrast, we accept that companies should lose immunity from liability when they 'promote' – or 'optimise' the presentation of – illegal content in the advertisement section of their platform as a result of commercial agreements.[159]

Recommendation 4: General monitoring of content must continue to be prohibited

ARTICLE 19 believes that governments must refrain from imposing an obligation of general monitoring of content by companies.

Although it may be argued that monitoring merely enables companies to detect potentially illegal or other problematic content, in practice, mere detection is almost always coupled with removal or other types of actions reducing the availability of such content. This is deeply problematic, given that content-monitoring technology is not nearly as advanced as is

sometimes suggested. In particular, hash-matching algorithms and natural language-processing tools are currently incapable of distinguishing content whose legality may vary depending on the context, such as news reporting or parody.[160] Vast amounts of legitimate content may therefore be removed. Moreover, these technologies interfere with users' privacy rights, as they require analysis of individuals' communications.

In addition, if a law were to make immunity from liability conditional on 'general monitoring' or the adoption of 'proactive measures' or 'best efforts' to tackle illegal content,[161] companies would inevitably err on the side of caution and remove content by default to avoid legal risks and enforcement costs. This could lead to platforms only allowing pre-screened speakers, or using their terms of service to prohibit controversial content,[162] and could also deter new market entrants from challenging incumbents.[163]

At the same time, ARTICLE 19 recognises that 'specific' monitoring and removal of videos or other images that contain incontrovertibly unlawful child sexual abuse images, i.e. the depiction of sexual activity (e.g. penetration) between a child and an adult, may be compatible with the rights to freedom of expression and privacy.[164] We do so given the gravity of the conduct at issue and the fact that this type of content can reliably be recognised as unlawful regardless of context. We do not, however, agree that such specific monitoring obligations should be applied to any other kind of content.[165]

Recommendation 5: Any regulatory framework must be strictly limited in scope

As noted earlier, ARTICLE 19 believes that any framework aiming to regulate platforms' content-moderation activities ought to be limited in its scope, including by reference to its subject matter, the entities it seeks to cover, and its geographical application. In particular, we make the following recommendations:

Regulation should focus on illegal rather than 'legal but harmful' content
ARTICLE 19 believes that any such framework should be limited to 'illegal' rather than 'harmful' content, for the simple reason that 'harmful' content is an inherently vague concept. This makes it difficult to enforce, prone to abuse, and open to challenge on legality grounds. In our view, legal content that is nonetheless prohibited under companies' community standards should be subject to oversight by independent multi-stakeholder entities, such as ARTICLE 19's proposed Social Media Councils.

If 'legal but harmful' content is included within the scope of legislation, contrary to our recommendations, then it should only impose transparency and due-process requirements for the enforcement of the company's community standards. The role of the regulator would therefore be limited to ensuring that companies' content-moderation systems are sufficiently transparent, and that users have clear and effective redress mechanisms available to them.

Private-messaging services and news organisations should be out of scope
Similarly, we believe that the scope of application of any regulatory framework should be limited so that below-the-line comments on newspaper websites and blogs are excluded. Similarly, journalistic content should, in principle, be excluded from scope, including when social-media platforms or other services (e.g. search engines) are hosting it. In our view, it would be widely inappropriate if measures aiming to regulate the practices of very large social-media platforms were to be used as a backdoor to regulating journalistic content.

Equally, messaging applications and other private channels of communication should be out of scope. In particular, regulators should not have the power to impose obligations on providers where such obligations would entail an unjustifiable interference with users' privacy rights, such as a weakening of end-to-end encryption or mandatory filters.

Measures should not have extraterritorial application
Finally, we believe that the implementation of measures under such a new regulatory framework should be geographically limited to the country mandating such measures, consistent with international principles of comity and the proportionality principle under international human rights law. In other words, no one country should be able to issue orders to remove or otherwise restrict content that may be lawful outside its borders.

Recommendation 6: Obligations under any regulatory scheme must be clearly defined

ARTICLE 19 believes that any obligations under a new regulatory scheme governing the activities of platforms and other tech companies must be clearly defined. Below, we set out the types of measures that could be included as part of such a framework and those that should not. In particular, we believe that a new regulatory framework could mandate the following:

Transparency obligations
In our view, transparency should be a basic requirement that pervades everything that companies do. In particular, it should apply to:

- **Distribution of content:** Social-media platforms and digital companies should provide essential information and explain to the public how their algorithms are used to present, rank, promote, or demote content. Content that is promoted should be clearly marked as such, whether the content is promoted by the company or by a third party for remuneration.[166] Companies should also explain how they target users with (unsolicited) promoted content, whether at their own initiative or on behalf of third parties as a paid service.[167]

- **Companies' terms of service and community standards:** Companies should publish community standards/terms of service that are easy to understand, and give 'case law' examples of how they are applied. They should publish information about the methods and internal processes for the elaboration of community rules, which should continue to include consultations with a broad range of actors, including civil society.[168]

- **Human and technological resources used to ensure compliance:** Companies should include detailed information about trusted-flagger schemes, including who is on the roster of trusted flaggers, how they have been selected, and any 'privileges' attached to that status. They should also publish information about how their algorithms operate to detect illegal or allegedly 'harmful' content under their community standards. In particular, this should include information about rates of false negatives/false positives and indicators, if any, to assess content that is likely to become viral, e.g. by reference to exposure to a wider audience.[169]

- **Decision-making**: Companies should notify affected parties of their decisions and give sufficiently detailed reasons for the actions they take against particular content or accounts. They should also provide clear information about any internal complaints mechanisms.

- **Transparency reports:** Companies should publish detailed information consistent with the Santa Clara Principles.[170] We note that it is particularly important not to limit statistical information to the removal of content but to also include data about the number of appeals processed and their outcome. Transparency reporting should also distinguish between content flagged by third parties (including whether they are public bodies or private entities), trusted flaggers (whether public bodies or private entities), and algorithms. Further information should also be provided about the different types of restrictions applied to content as part of content-moderation processes, such as demonetisation or downgrading; for every restriction, the company should give information about the rules the decision was based on, and, where available, the outcome of any appeals.

  More generally, we note that any transparency reporting requirements should aim to provide a far more qualitative analysis of content-moderation decisions. The metric of success in addressing illegal content must not be tied to content-removal rates, as this encourages over-removal. Equally, transparency reporting should not be limited to information submitted by companies but should also include information submitted by relevant government agencies. The above is without prejudice to any measures that may be applicable under consumer law.[171]

- **Transparency audits:** Companies should give greater access to datasets to regulators and vetted independent researchers – whether academics, journalists, or otherwise – to enable them to verify that the company's systems and algorithms are operating as the company says they do. In particular, auditors should be given access to data about: (1) companies' content-moderation programmes; (2) how companies order, rank, prioritise, recommend, or otherwise personalise content; and (3) how this applies to political advertising.[172] While regulators could be given access to sensitive and commercial data, vetted third parties could be given access to anonymised datasets. These audits of platforms' operations should take place on a regular basis.

- **Archives of digital and political advertising:** New rules should ban 'data opacity' for political ads and ask platforms for clear enforcement mechanisms for violation of their policies. Platforms should not be placed in the role of refereeing or mitigating aggressive political discourse and 'disinformation.' In addition, we call for enhanced transparency concerning all political ad spending from relevant stakeholders, including political parties, tech companies, and third-party advertisers. Political ads should be clearly distinguishable from editorial content, including news, whatever their form and including online, and clearly labelled with information about who paid for them. Furthermore, we support the use of digital ad databases to keep and publish all regulated ads, the amount of money spent on advertising, and the name of the person who authorised the ad, which should be accessible in a format that allows for bulk retrieval by researchers and policy makers.

Internal due-process obligations
Without prejudice to Recommendation 9, ARTICLE 19 believes that any regulatory framework regulating the activities of dominant platforms should include a requirement to put in place:

- **Clear notice and action rules,** in line with the Manila Principles on Intermediary Liability.

- **Internal redress mechanisms** to deal with complaints about restrictions on the exercise of the right to freedom of expression, such as the wrongful removal of content or the wrongful application of labels that would suggest that a news source is untrustworthy. Conversely, appeals mechanisms should also be able to address a company's refusal to remove content that is arguably in breach of the company's community standards. In all cases, internal complaints mechanisms should respect due-process safeguards.[173]

- **Obligation to promote media diversity:** Given the risks of overly personalised content on social-media platforms, very large social-media companies should be required to take steps to ensure a sufficient degree of media diversity,[174] as well as balanced coverage during elections, on their service.[175] In particular, they should provide sufficient information to explain how newsfeeds and the material they promote is selected. They should also provide users with viable alternatives to choose from that are not based on profiling. In certain limited circumstances, such as during elections, they could be required to carry messages from public-service media to ensure that the widest possible segment of the population has the basic information they need to participate meaningfully in elections.

Refrain from imposing certain obligations
By contrast, ARTICLE 19 believes that any such regulatory scheme should not include the following – non-exhaustive – types of obligations:

- **A broad and undefined 'duty of care' to prevent an equally undefined notion of 'harm':** In our view, such notions would be unlikely to pass the legality test under international human rights law. In practice, they would both create legal uncertainty and give largely unfettered powers to regulatory authorities, which would be deeply problematic for freedom of expression.

- **A general obligation to monitor content:** As noted above, there should be no obligation of general monitoring of content, or measures that are substantially equivalent to it, such as mandating 'best efforts' or 'proactive measures' to tackle illegal content. Equally, such a framework should refrain from 'nudging' companies towards the adoption of such measures by framing them as purely voluntary or simply 'recommended', when, in reality, failure to adopt them could lead to heavy sanctions.

- **Unduly short timeframes:** Companies should not be required to remove content within unduly short timeframes, particularly when the content at issue may give rise to difficult questions of interpretation, such as 'hate speech' or 'terrorist' content. Short removal timeframes do not incentivise companies to review notices with sufficient care. As such, they promote the wrongful removal of content and fail to protect freedom of expression. Moreover, removals within a short timeframe can incentivise companies to allocate resources to the removal of notices regardless of their severity or to focus on content simply because it has been posted in the last 24 hours, rather than older content that may well be more deserving of attention.[176]

- **Setting up compliance targets:** Equally, legislation or regulators should not impose numerical compliance targets that could have the effect of encouraging companies to expand the definition of content they disallow on their platform to boost their compliance rate.[177] In other words, numerical targets would encourage the removal of ever-greater amounts of legitimate content. We further note that, insofar as lawmakers may be considering various metrics and thresholds to ensure compliance, they should consider the extent to which society can be expected to tolerate a degree of risk of harm online, as it does in the offline world.[178]

- **Obligation to cooperate or report illegal content:** Vague obligations to cooperate are problematic because they could involve serious interferences with users' rights, such as access to user data by law enforcement without sufficient safeguards. At the same time, being vague makes it arguable for companies that they have cooperated in other less intrusive ways. In short, such obligations are likely to be difficult to enforce; as such, it is unclear whether they are necessary. Obligations to report illegal content would likely give companies a strong incentive to focus on notices they receive of allegedly illegal content, regardless of its severity, and report it to law enforcement. They could also disincentive companies to invest in automated tools to detect illegal content if they could be fixed with knowledge of illegality, or found in breach of their obligations, for failing to report all the potentially illegal content they identify automatically on their networks. Both these outcomes would be undesirable, and would likely have a negative impact on freedom of expression, since vast amounts of legitimate content would be reported.

- **Must-carry obligations:** While we believe that 'must-carry obligations' – i.e. a requirement for platforms to publish lawful content that is otherwise in breach of their terms of service – may be imposed in very limited circumstances, in our view, it would be inappropriate to impose more general must-carry obligations. In particular, must-carry obligations would interfere with both the free-speech rights and the right to property of social-media platforms.[179] The right to free expression of platforms must imply that they should be free to allow or disallow e.g. far-right conspiracy theories on their site, even though such theories might be lawful in some countries. As a matter of proportionality, users who wish to promote such theories would still have other channels available to them to express those views. In practice, general must-carry obligations would also almost certainly undermine content moderation, which would be undesirable.

Recommendation 7: Any regulator must be independent and accountable in both law and practice

ARTICLE 19 believes that, for any online content-regulatory scheme to have any kind of legitimacy, it must be overseen by an independent regulator; that is, a regulator free from political or commercial interference.

A regulator's independence and institutional autonomy must be guaranteed and protected by law, including through:

- Clear statement of overall platform and online content-regulation policy;

- Clearly laying out the powers and responsibilities of the regulator;

- Rules of membership; and

- Funding arrangements and accountability to the public through a multi-party body. The government should be kept at arms-length and not be involved in any of those funding arrangements and accountability bodies.

More specifically, for freedom of expression to be protected by a regulator, the law setting it up should contain:

- Overarching provisions stressing the importance of protecting freedom of expression, including expression that may shock, offend, or disturb;

- Provisions making clear that the mission of any regulator in this area is to protect human rights, including freedom of expression;

- Provisions requiring any regulator to audit content-removal decisions and consider the extent to which companies over-remove or over-restrict content, whether upon request or of their own accord; and

- Provisions making clear that companies should not be penalised for failing to remove lawful content.

Any regulator tasked with overseeing the operations of a broad range of providers of digital services should ensure cooperation with other relevant regulators, such as data-protection, consumer-protection, and competition authorities.

Finally, any regulator must be accountable to the public, including through transparency obligations about its activities and annual reports laid before parliament.

Recommendation 8: Any regulatory framework must be proportionate

ARTICLE 19 believes that, for any regulatory framework to comply with international standards on freedom of expression, it must be strictly proportionate to the aim pursued:

- **Tiered approach:** Governments should be extremely cautious about adopting measures that are meant to hold large social-media companies to account but would ultimately impose an undue burden on other, smaller, services. As such, we believe that a tiered approach in this area would be necessary. In other words, large social-media platforms could be made subject to more stringent obligations than smaller players. To assess the dominance of a platform, regard could be had to the following factors: (1) the number of its users; (2) its annual global turnover; and (3) its market power. Non-profits, such as Wikipedia, should be exempt and continue to operate under a broad immunity from liability framework.

  At the same time, we highly recommend that any proposed measures should be the subject of rigorous impact assessments, including possible anti-competitive outcomes. Large social-media companies are likely to be able to adapt to any demands placed upon them. Such demands could ultimately lead to a perception that they are 'safer'; this would almost certainly give them an advantage over smaller entrants, which would not be able to engage in the same kind of content-moderation exercise as the incumbents.

- **Evaluating systemic failures:** For regulation in this area to be sustainable and proportionate, companies should not be assessed because they have failed to remove a single piece of content or only published a single dataset. Instead, a regulator should evaluate whether they have failed to comply with their obligations under the law on a systemic basis. The threshold for systemic failures should be defined by law by reference to clear criteria that should enable a holistic assessment rather than a purely numerical one. For instance, the law should not sanction companies because they have failed to remove a given quantity or percentage of content flagged as either illegal or harmful. Rather, it should contain an overall assessment of the measures they have adopted to mitigate risks to human rights.

- **Proportionate sanctions:** Failure to comply with the obligations outlined above should be meted out with proportionate sanctions. While this may include significant fines, these should not be set so high as to provide a disincentive to protect freedom of expression. In our view, 4% of global turnover is likely to be too high for freedom of expression to be protected. Equally, criminal sanctions imposed on chief executives for failure to comply with these obligations could have a chilling effect on freedom of expression. Faced with the prospect of several years in prison, company executives would almost certainly adopt policies that would favour greater removal or other types of restrictions on content. As such, governments should refrain from adopting such criminal sanctions.

Recommendation 9: Any regulatory framework must provide access to effective remedies

Beyond internal complaints mechanisms, ARTICLE 19 believes that governments should ensure that individuals have access to judicial remedies to challenge wrongful removal of their content by social-media platforms on the basis of their terms of service. Such remedies should include not only access to the courts but also alternative dispute-resolution mechanisms, such as e-courts or an ombudsman.[180] In practice, governments should develop proposals for funding such mechanisms, including through e.g. a levy on social-media platforms.

This should be without prejudice to self-regulatory schemes, such as Social Media Councils, that would (among other things) enable users to challenge social-media platforms' content-moderation decisions by reference to an agreed set of principles, such as a 'Charter of Users' Rights'.

Recommendation 10: Large platforms should be required to unbundle their hosting and content-curation functions and ensure they are interoperable with other services

ARTICLE 19 believes that, to address the excessive market power of a handful of social-media platforms, content curation should be decentralised.

Regulators could mandate large platforms to separate their hosting and content-curation functions to allow third parties to access their platform (in practice, their API) to provide content-curation services to users.[181]

This form of functional separation would not impede the large social-media platforms from offering content curation to their users. However, users would decide whether to opt in. In

other words, when creating a profile on Facebook, for example, the user should be asked to select a content-curation provider, and Facebook could remain one of the options to select. Ideally, and to avoid further lock-in, users should remain free to change their choice at any time through the platform's settings. In our view, these kinds of solutions should be further explored to enable users to take back control, ensure healthy competition and innovation in social-media markets, and return to the promise of a diverse and decentralised Internet.[182]

At the very least, we believe that social-media platforms should give users greater content-curation options, both in terms of the type of content they would like to view more of and according to what criteria, e.g. in chronological order.

ENDNOTES

[1] John Perry Barlow, A Declaration of the Independence of Cyberspace, Electronic Frontier Foundation, 8 February 1996.

[2] 'Hosts' are typically those that rent web-server space to enable their customers to set up their own websites. However, the term 'host' has also taken on a more general meaning, i.e. any person or company who controls a website or a webpage that allows third parties to upload or post material. For this reason, social-media platforms and video- and photo-sharing services are usually referred to as 'hosts'. 'Hosting' is then a role characterised by the absence of editorial intervention in content; *c.f.* ARTICLE 19, Internet intermediaries: Dilemma of liability, 2013, p.6.

[3] According to recent estimates, there are currently 2.5 billion monthly active Facebook users and 330 million Twitter users, and 1 billion hours of video are watched daily on YouTube. See e.g. Statistics Portal, Leading countries based on Facebook audience size; Statistics Portal, Number of monthly active Twitter users worldwide from 1st quarter 2010 to 1st quarter 2019; YouTube for Press, YouTube in numbers.

[4] See e.g. Twitter, Permanent suspension of @realDonaldTrump, 8 January 2021; Guy Rosen Monika Bickert, Our response to the violence in Washington, Facebook, 6 January 2021.

[5] K. Wagner and P. Martin, Twitter locks out Chinese Embassy in U.S. over post on Uighurs, Bloomberg, 20 January 2021.

[6] See e.g. Section 230 of the Communications Decency Act 1996 (47 USC. § 230) in the USA, or EU Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (Directive on electronic commerce) ('E-Commerce Directive').

[7] See e.g. the UK's Online Harms White Paper, 2019, and Draft Online Safety Bill, 2021; the EU's Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and amending Directive 20000/31/EC, 2020; Ireland's Online Safety and Media Regulation Bill, 2020; the Australian Online Safety Bill, 2021; India's Draft Intermediaries Guidelines (Amendment) Rules, 2018.

[8] See ARTICLE 19, #MissingVoices campaign.

[9] See e.g. Online Harms White Paper: Full government response to the consultation, 15 December 2020; or the French Draft Bill on Countering Online Hatred (so call Loi Avia, or Avia Bill), which the French Constitutional Council (Conseil d'État) subsequently declared unconstitutional.

[10] Dilemma of liability, *op.cit.*

[11] ARTICLE 19, Sidestepping rights: Regulating speech by contract, 2018.

[12] *Ibid.*

[13] The policy also refers to several legislative proposals to platform governance that were recently put forward in some countries, as these might be used elsewhere as inspiration for the adoption of similar legislation, in particular the Draft Online Safety Bill, 2021, and the EU proposal for the Digital Services Act (see the European Commission, The Digital Services Act package). It is our understanding that these proposals might have been amended or changed in the interim, and this policy cites them in their form at the time of publication (December 2021).

[14] ARTICLE 19, Taming big tech: Protecting freedom of expression through the unbundling of services, open markets, competition, and users' empowerment.

[15] ARTICLE 19, policy paper on media diversity online, forthcoming.

[16] ARTICLE 19, policy paper on must-carry obligations, forthcoming.

[17] These include, for instance, the issues of advertising, targeted ads, and the business models of companies and freedom of expression and information.

[18] Dilemma of liability, *op.cit.*

[19] See e.g. L. Lovdahl Gormsen and J.T. Llanos, Facebook's anticompetitive lean in strategies, *SSRN*, 6 June 2019.

[20] *C.f.* e.g. Communication from the Commission to the European Parliament, The Council, the European Economic and Social Committee and the Committee of the Regions, Online Platforms and the Digital Single Market Opportunities and Challenges for Europe, COM (2016) 288 final, 25 May 2016.

[21] See also Sidestepping rights, *op.cit.*, p. 36.

[22] See e.g. Dilemma of liability, *op.cit.*

[23] At common law, see e.g. D. Rolph, Liability in the Internet age, International Forum for Responsible Media (IFFRM) Blog, 17 February 2011.

[24] 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).

[25] UNGA, International Covenant on Civil and Political Rights (ICCPR), 16 December 1966, UN Treaty Series, vol. 999, p.171.

[26] 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.

[27] Human Rights Committee (HR Committee), General Comment No. 34 on Article 19: Freedoms of opinion and expression, CCPR/C/GC/34, 12 September 2011, paras 12, 17, and 39.

[28] 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 Organization of American States Special Rapporteur on freedom of expression (OAS Special Rapporteur on FoE), and the African Commission on Human and Peoples' Rights (ACHPR) Special Rapporteur on freedom of expression and access to information, June 2011.

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

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

[31] General Comment No. 34, *op.cit.*, para 43.

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

[33] Special Rapporteur on FoE, Report of 16 May 2011, A/HRC/17/27, para 43.

[34] Special Rapporteur on FoE, Report of 6 April 2018, A/HRC/38/35, para 66.

[35] *Ibid.*

[36] *Ibid.*

[37] Special Rapporteur on FOE, Report of 13 April 2021, A/HRC/47/25, para 91.

[38] Guiding Principles on Business and Human Rights: Implementing the UN 'Protect, Respect and Remedy' Framework (the Guiding Principles), 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.

[39] *Ibid.,* Principle 15.

[40] *Ibid.,* Report of the Special Rapporteur on FoE, 6 April 2018, A/HRC/38/35, paras 48 and 7; OAS Special Rapporteur on FOE, Standards for a free, open and inclusive Internet, 15 March 2017, paras 111–112.

[41] *Ibid.*, the 2018 Report of the Special Rapporteur on FoE, paras 45–46; the 2013 OAS Special Rapporteur on report, para 113. See also OAS Special Rapporteur on FOE, Standards for a free, open and inclusive Internet, 15 March 2017, para 99.

[42] *Ibid*., Special Rapporteur on FoE, paras 47 and 76; the 2013 OAS Special Rapporteur on FoE, para 115.

[43] The April 2018 Report of the Special Rapporteur on FoE, *op. cit*., para 40 – 44.

[44] *Ibid.*, paras 70–72.

[45] Council of Europe Commissioner for Human Rights, The rule of law on the Internet and in the wider digital world, CommDH/IssuePaper (2014) 1, 8 December 2014.

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

[47] 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. These recommendations were further echoed in the Committee of Ministers, Guide to human rights for Internet users, Recommendation CM/Rec(2014)6 and explanatory memorandum, p.4.

[48] Recommendation CM/Rec (2018) 2 of the Committee of Ministers to Member States on the roles and responsibilities of Internet intermediaries, adopted by the Committee of Ministers on 7 March 2018.

[49] See 2013 OAS Report, *op. cit.* and 2016 OAS Report, *op.cit*.

[50] *Ibid*.

[51] See the Declaration of Principles on Freedom of Expression and Access to Information in Africa, 2019, adopted by the ACHPR at its 65th ordinary session in Banjul, the Gambia.

[52] Declaration of Principles on Freedom of Expression and Access to Information in Africa. See also The African Declaration of Internet Rights and Freedoms, 2014, developed as a civil society initiative.

[53] The Manila Principles on Intermediary Liability (the Manila Principles), March 2015.

[54] *Ibid*., Principle 4.

[55] *Ibid*., Principle 5(c).

[56] Ranking Digital Rights, Corporate Accountability Index: 2015 Research Indicators, June 2015.

[57] The 2016 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 FoE, and the ACHPR Special Rapporteur on freedom of expression and access to information, 4 May 2016.

[58] The 2017 Joint Declaration on Freedom of Expression and 'Fake News', Disinformation and Propaganda, the UN Special Rapporteur on FoE, the OSCE Representative on freedom of the media, the OAS Special Rapporteur on FoE, and the ACHPR Special Rapporteur on freedom of expression and access to information, 3 March 2017.

[59] 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, joint press release, 8 March 2017.

[60] The Report of the Special Rapporteur on FoE, A/74/486, 9 October 2019, para 57(d).

[61] 2016 Joint Declaration, *op.cit*., para 2 (e).

[62] 2017 Joint Declaration, *op.cit*., para 4 (b).

[63] *Ibid*., para 58 (b).

[64] The Report of the Special Rapporteur on FOE, A/73/348, 29 August 2018.

[65] *Ibid*., para 64.

[66] *Ibid.*

[67] *Ibid*., para 66.

[68] *Ibid.*

[69] *Ibid*., para 70.

[70] See Article 17 of the ICCPR, *op.cit*. The UN Special Rapporteur on the 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.

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

[72] *Ibid*., para 84.

[73] 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.

[74] *Ibid.* See also the Special Rapporteur on FoE, Encryption and anonymity follow-up report, Research Paper 1/2018, June 2018.

[75] For example, in the third quarter of 2020, Facebook registered 2.5 billion monthly active users worldwide – almost the size of the populations of China and India combined. In 2017, Facebook's annual revenues were superior to the GDP of Serbia; see F. Belinchón and Q. Moynihan, 25 giant companies that are bigger than entire countries, *Business Insider España*, 25 July 2018. Similarly, in December 2019, Google had over 90% market share of search services worldwide; see Statcounter, Search engine market share worldwide, April 2019–April 2020. Twitter, while smaller, still has over 125 million daily active users, i.e. about 2.5 times the population of Kenya.

[76] See e.g. the UK's proposals on Online Harms/Online Safety Bill, *op. cit.,* the now defunct French Loi Avia Bill, *op. cit.,* or the Austrian Draft Communication Platforms' Act, October 2020.

[77] For instance, the French Loi Avia Bill, which the French Constitutional Council declared unconstitutional in June 2020, provided that the government sets thresholds over which companies are expected to fulfil certain obligations, such as takedown of illegal hate speech content within 24 hours (see note 7). Similarly, in Germany, a draft law on media diversity lays down a number of obligations on 'video-platforms' with more than 1 million users in Germany per month; the Proposal on issues of radio interface, platform control and intermediate media, State-specific regulations for approval, platform regulation and intermediaries (so-called *Medienstaatsvertrag*), 2018.

[78] For instance, the 2019 EU Copyright Directive provides that online content-sharing service providers (OCSSPs) will lose immunity from liability unless they fulfil a number of strict conditions. However, compliance with these new obligations under the Directive must take into account 'the type, the audience and the size of the service', in accordance with the proportionality principle. Companies that have provided services for fewer than three years in the European Union and have a turnover below EUR 10 million are subject to less stringent conditions so as to benefit from immunity from liability; see Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC. By contrast, the 2019 EU Copyright Directive provides that OCSSPs with an average number of monthly unique visitors exceeding 5 million are subject to additional requirements.

[79] See e.g. Ofcom, Online market failures and harms: An economic perspective on the challenges and opportunities in regulating online services, 28 October 2019; Australian Competition and Consumer Commission, Digital Platforms Inquiry: Final report, 26 July 2019; and the French and German competition authorities, Algorithms and competition, joint report, November 2019.

[80] For instance, Cloudflare decided to terminate the provision of its services to Daily Stormer – a far-right, neo-Nazi, white-supremacist message-board website – in 2017, following a fatal car attack against anti-far-right protesters in Charlottesville. In 2019, Cloudflare again terminated the provision of its services to 8chan, a far-right imageboard, after the perpetrator of the El Paso shooting allegedly posted his manifesto on that website; see M. Prince, Why we terminated Daily Stormer, Cloudflare, 16 August 2017; M. Prince, Terminating service for 8Chan, Cloudflare, 5 August 2019.

[81] See e.g. Amnesty International, Surveillance giants: How the business model of Google and Facebook threatens human rights, POL 30/1404/2019, November 2019; Privacy International, Privacy International's response to the open consultation on the Online Harms White Paper, July 2019.

[82] *Ibid.*

[83] See e.g. K. Iwańska, To track or not to track: Towards privacy-friendly and sustainable online advertising, Panoptykon Foundation, November 2020.

[84] *Ibid.*

[85] *Ibid.*

[86] See e.g. Access Now, Raising the alarm: Online tracking harms human rights, 14 December 2020

[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 that should be protected. Given the complexity around this term, ARTICLE 19 refers to it as 'hate speech' in this and other policy documents.

[88] For instance, in the UK, there was a public outcry after Molly Russell took her own life and graphic posts about suicide and self-harm were found on her Instagram account; see e.g. BBC News, Molly Russell: Did her death change social media?, 27 October 2019.

[89] For example, in the area of disinformation, Facebook's policy has significantly changed over time, from focusing on 'inauthentic behaviour' (i.e. use of fake accounts) to explicitly removing or reducing the spread of 'false news' or 'manipulated media' (e.g. deep fakes); see e.g. ARTICLE 19, Facebook Community Standards: Analysis against international standards on freedom of expression, 30 July 2018. Twitter, for its part, only removes limited misleading information in the context of elections but continues to tackle 'platform manipulation' and spam; see Twitter rules on election integrity, 2021, and Twitter, Platform manipulation and spam policy, September 2020. As for 'hate speech', social-media companies have had to take steps to address the serious criticisms directed at them by various human rights groups, journalists, lawmakers, and institutions such as the UN. For instance, Amnesty International denounced Twitter for failing to protect women's rights online and creating a toxic environment, driving women away from public discourse; see Amnesty International, Toxic Twitter: A toxic place for women, 15 March 2018. YouTube came under fire for failing to take action against conservative YouTube host Steven Crowder after he consistently harassed Vox journalist Carlos Maza, including by using racist and homophobic slurs in YouTube videos about him over a period of two years; see E. Stuart, "We don't want to be knee-jerk": YouTube responds to Vox on its harassment policies, Vox, 10 June 2019. The UN investigators cited

the 'role' of Facebook in spreading 'hate speech' – and therefore facilitating the possible genocide of the Rohingya people in Myanmar; see Report to the Human Rights Council of the independent international fact-finding mission on Myanmar, 12 September 2018, A/HRC/39/64, para 74.

[90] For example, within the EU, these include the Code of Practice on Disinformation, 17 June 2019, and the EU Code of Conduct on Countering Illegal Hate Speech, 2018. On the global level, the Global Internet Counter-Terrorism Forum (GIFCT), an initiative that Facebook, Microsoft, Twitter, and YouTube launched in 2017, enables its members to access a database of hashes that members can use to decide whether to remove content on their own service and on the basis of their own community standards. Under the Christchurch Call, a global initiative the New Zealand government launched following the live-streaming of terrorist attacks in Christchurch, online service providers commit to, inter alia, 'take transparent, specific measures seeking to prevent the upload of terrorist and violent extremist content and to prevent its dissemination on social media and similar content-sharing services, including its immediate and permanent removal'.

[91] Some States have already taken the view that these measures were insufficient, and have adopted legislation banning disinformation, such as France and Singapore.

[92] In the EU, the Digital Services Act is also said to be necessary in order to ensure the efficiency of the single market; *op.cit.* In the UK, a single regulator to deal with the various problems posed by content online has the attraction of apparent simplicity; see Online Harms/Online Safety Bill proposal, *op.cit*.

[93] See French Secretary of State for Digital Affairs, Creating a French framework to make social media platforms more accountable: Acting in France with a European vision, Mission Report, version 1.1, May 2019, p.11; ARTICLE 19, France: ARTICLE 19 comments on French interim report for social media regulation, 19 June 2019.

[94] For instance, in the US, Senator Hawley has sought to introduce legislation to amend Section 230 of the Communications Decency Act (CDA) so that platforms would lose their immunity from liability unless an external audit convincingly demonstrates that their algorithms and content-removal practices are politically neutral; see Senator Hawley introduces legislation to amend Section 230 immunity for Big Tech companies, 19 June 2019. While the draft law has not been adopted, the debate around Section 230 still rages on in the US. For example, in January 2020, Joe Biden said he wanted to revoke it; see e.g. Makena Kelly, Joe Biden wants to revoke Section 230, *The Verge*, 17 January 2020. In Germany, the Broadcasting Authority has put forward a proposal that would impose diversity obligations on video and social-media platforms for the first time; see N. Helberger, P. Leerssen, and M. Van Drunen, Germany proposes Europe's first diversity rules for social media platforms, LSE Blog, 29 May 2019.

[95] See e.g. E. M. Mazzoli and D. Tambini, Prioritisation uncovered: The discoverability of public interest content online, Council of Europe Study DGI, 2020, 19, p.40ff.

[96] For more details on the challenges raised by the distribution of news by big social media platforms, see P.F. Docquir & M.L. Stasi, The Decline of Media Diversity – and How we can Save it, 21 January 2020.

[97] The Guardian, How social media filter bubbles and elections and algorithms influence the election, 22 May 2017.

[98] The New York Times, Conservatives accuse Facebook of Political Bias, 10 May 2016.

[99] See for instance, The Guardian, Watchdog cracks down on tech firms that fail to protect children, 22 January 2020.

[100] See, e.g. D. Keller & J. von Hoboken, Design Principles for Intermediary Liability Laws, Transatlantic High Level Working Group on Content Moderation, October 2019.

[101] See ARTICLE 19, Germany: The Act to Improve Enforcement of the Law in Social Networks, August 2017; ARTICLE 19, France: Analysis of draft hate speech bill, 3 July 2019.

[102] For instance, initiatives such as the Global Internet Forum on Countering Terrorism (GIFCT) have remained incredibly opaque. Little data is available about the kind of content that is removed on the basis of the GIFCT hash-database; see Transparency page on the GIFCT website. In Germany, opinion remains divided on the effectiveness of the NetzDG law, which is largely aimed at tackling 'hate speech' online. Experts have noted the difficulties in providing an objective assessment of the law given the lack of uniformity of transparency reporting by social media companies. However, there is a degree of consensus that the new law has contributed to the faster removal of content, though its actual impact on 'hate speech' remains unclear and hard to prove empirically: see e.g. A. Held, Germany is amending its online speech act NetzDG…but not only that, 6 April 2020.

[103] In France, for instance, a group of LGBTI activists recently wrote an open letter in the newspaper *Libération* to state that the French draft online hate speech (Avia) law would be counter-productive and lead to an increased number of removals of content produced by LGBT communities. The group demanded more resources for the justice system to prosecute the actual perpetrators of hate crimes, including the offences of racist insults, among others; see Libération, Feminist, LGBTI and antiracist: we do not want cyber-hate law, Tribune, 21 January 2020.

[104] These include, for example NetzDG in Germany and the Loi Avia Bill.

[105] See e.g. Facebook's recent updates; YouTube, Our ongoing work to tackle hate, 5 June 2019.

[106] *Ibid.* See also Wired, Twitter and Instagram unveil new ways to combat hate – again, 7 November 2019.

[107] See e.g. Google, Self-reporting on Google News Initiative, one year in, 20 March 2019.

[108] See e.g. EU Code of Conduct on Countering Illegal Hate Speech, Factsheet, February 2019.

[109] The Verge, Twitter rolls out 'hide replies' to let you tame toxic discussions, 19 September 2019.

[110] For the shortfalls of these appeals mechanisms, see ARTICLE 19, Sidestepping Rights, *op. cit.*, and , #Missing Voices Campaign.

[111] See, e.g. Facebook, An Independent Assessment of the Human Rights Impact of Facebook in Myanmar, 5 November 2018; and Facebook, A Second Update on Our Civil Rights Audit, 30 June 2019.

[112] See, e.g. Forbes, Facebook is Hiring 3,000 Moderators In Push To Curb Violent Videos, 03 May 2017; Financial Times, Facebook steps up efforts to combat hate speech in Myanmar, 16 August 2018; or Telegraph, Facebook to Hire 1000 new workers in the UK to help fight toxic content, 21 January 2020.

[113] For example, in the EU, the European Commission noted in its first January 2019 monitoring report overseeing the implementation by social media companies of their commitments under the EU Code of Practice on Disinformation that it was encouraged by the policies developed by companies to tackle disinformation but was deeply concerned about the lack of specific benchmarks to measure progress; see European Commission, First monthly intermediate results of the EU Code of Practice against disinformation, 28 February 2019. In its last monitoring report in May 2019, the European Commission noted the progress achieved in improving the **transparency of political advertising** and public disclosure of such ads, but added that more had to be done to tackle disinformation effectively; see European Commission, Last intermediate results of the EU Code of Practice against disinformation, 14 June 2019.

[114] See, for instance, King's College London Centre for the Study of Media, Communication and Power, Submission to the Culture Media and Sport Select Committee's Inquiry into Fake News, 2017.

[115] This is without prejudice to ARTICLE 19's more detailed position on political advertising and disinformation, forthcoming.

[116] See e. g, B. Wagner and L. Kuklis, Disinformation, data verification and social media, 16 January 2020.

[117] For instance, the European Commission noted in its fourth evaluation of the EU Code of Conduct on Countering Illegal Hate Speech that this initiative had delivered successful results. In particular, it highlighted that IT companies were now able to assess 89% of flagged content within 24 hours, 72% of the content deemed to be illegal 'hate speech' was removed, compared to 40% and 28% respectively when the Code was first launched in 2016. See European Commission, *op. cit.*; see also European Commission, Countering illegal speech online: EU Code of Conduct ensures swift response, 04 February 2019. Similarly, the French found that 'the speed of deployment and progress made during the last 12 months by an operator such as Facebook show the benefits of capitalising on this self-regulatory approach already being used by the platforms, by expanding and legitimising it', and that 'the self-regulatory capacity observed at these operators providing a social network makes it possible to position them as key elements in the solution to social cohesion issues raised by the presence of certain content on these platforms'; see French Secretary of State for Digital Affairs, Creating a French framework to make social media platforms more accountable: Acting in France with a European vision, May 2019, p. 11.

[118] For more information about ARTICLE 19's social media council, see ARTICLE 19, Self-regulation and 'hate speech' on social media platforms, 2 March 2018; ARTICLE 19, Social Media Councils: Consultation, 11 June 2019.

[119] See Explanatory Memorandum to the EU proposals for a Digital Services Act, *op. cit.,* p. 1.

[120] See European Commission, Leaked note on a Digital Services Act, Netztpolitik, June 2019.

[121] General Comment No. 34, *op.cit.*, para 43.

[122] See e.g. R. Fletcher, Polarisation in the news media, Digital News Report, Reuters Institute and University of Oxford, 2017.

[123] See ARTICLE 19 policy on political advertising, forthcoming.

[124] For an overview of selective legislative proposals, see ARTICLE 19, Kyrgyzstan: Platform regulation laws and freedom of expression, January 2021.

[125] See, for example, the EU draft proposals for a Digital Services Act, *op. cit,;* the Austrian Draft Communication Platforms Act, *op. cit.;* or the now defunct Loi Avia Bill, *op. cit*. This is also one of the positive features of the German NetzDG law, see e.g. B. Wagner, K. Rozgonyi, M.T. Sekwenz, J. Cobbe and J. Singh, Regulating Transparency?: Facebook, Twitter and the German Network Enforcement Act, January 2020.

[126] See, for example, the EU's proposals for a Digital Services Act, *op. cit.* or the Austrian Draft Communication Platforms Act 2020, *op. cit.*

[127] *Ibid.*

[128] A 'Good Samaritan' clause refers to a provision that ensures that platforms or other hosting content providers are not penalised by losing immunity from liability if they take steps to moderate content of their own initiative. See e.g. J. Barata, Positive Intent Protections: Incorporating a Good Samaritan principle in the EU Digital Services Act, The Center for Democracy and Technology, July 2020.

[129] See e.g. the EU proposal for a Digital Services Act, *op. cit.* or the now defunct draft French Loi Avia Bill, *op. cit.* It is also a significant feature of the German NetzDG law, see e.g. ARTICLE 19's analysis of the Draft NetzDG law, 2017.

[130] See, for example, the EU proposal for a Digital Services Act, the UK Online Harms/Online Safety Bill proposal, *op.cit.*, Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (EU Copyright Directive 2019), the now defunct Draft Avia law on online hate speech, *op. cit,* or the more recent Austria Draft Communication Platforms Act 2020, *op. cit.* It is also a significant aspect of the German NetzDG law, *op. cit.*

[131] See the proposals in the UK, Ireland, Australia and in the leaked EU memo about the Digital Services Act, *op.cit.*

[132] *Ibid.,* the proposal in Ireland.

[133] See the UK Online Harms/Online Safety Bill proposals, *op.cit.*

[134] See, e.g. NetzDG in Germany or Loi Avia Bill in France, *op.cit.*

[135] See, e.g. UK Online Harms/Online Safety Bill proposals, *op.cit.*

[136] For instance, under the Irish proposal, the Online Safety Commissioner would have discretion to designate "any online service that allows users to share, spread or access content that other users have made available," *op.cit*

[137] See the French Facebook mission report proposal, *op.cit.*

[138] See the UK Online Harms/Online Safety Bill proposals, *op.cit.*

[139] See the UK government, Online Harms White Paper - Initial consultation response, February 2020.

[140] See ARTICLE 19, Response to the Consultations on the White Paper on Online Harms, June 2019.

[141] *C.f.* proposal in India, *op.cit.*

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

[143] *Ibid.,* Principle IV.

[144] See, e.g. proposals in France, Ireland, the UK and Australia, *op.cit.*

[145] See, e.g. proposals in France or Australia, *op.cit.*

[146] See, e.g. now defunct Loi Avia Bill in France or NetzDG in Germany, *op.cit.*

[147] See, e.g. now defunct Loi Avia Bill in France, *op.cit*.

[148] For instance, ARTICLE 19 was part of a Joint International Press Freedom Mission to Hungary that concluded that the Media Council is not independent with all its members being nominated by the Fidesz party; see ARTCLE 19, Hungary: Conclusions of the Joint International Press Freedom Mission, 3 December 2019. See also ARTICLE 19 and Polish Helsinki Foundation for Human Rights, Poland: independence of public service media, 31 January 2017.

[149] For instance, the NetzDG does not define what amounts to 'systemic' failures to comply with obligations under the law. We are further concerned by proposed regulators' powers to order the blocking of websites and the criminalisation of senior management for their failure to comply with a vague 'duty of care'; *op.cit.*

[150] For example, the Indian police raided Twitter's offices in Delhi, India after the company labelled the tweets of members of the ruling party as 'manipulated' media, see Buzzfeed, Police In Delhi Have Descended On Twitter Headquarters In The Country, 24 May 2021.

[151] For example, regulatory approaches proposed in Europe may not be easily transposed or appropriate elsewhere.

[152] See: Dilemma of liability *op. cit.* and Sidestepping Rights, *op.cit.*

[153] *C.f.* the discussions on Digital Services Act in the EU, *op.cit.*

[154] This means that companies such as Cloudflare should not be penalised for hosting websites such as 8chan or DailyStormer; *op.cit.*

[155] Dilemma of liability, *op.cit*.

[156] See e.g. the Internet Watch Foundation in the UK or SaferNet in Brazil.

[157] The Manila Principles, *op.cit*

[158] This is also consistent with the case law of the majority of EU Member States; see e.g. European Commission, Overview of the notice and action procedures in EU member states, SMART 2016/0039, July 2018.

[159] See *L'Oreal v eBay*, C-324/09, 12 July 2011.

[160] For a detailed discussion of content moderation technologies, see Center for Democracy and Technology, Mixed Messages: the limits of automated social media content analysis, November 2017.

[161] *C.f.* the EU Copyright Directive in the Digital Single Market.

[162] D. Keller & J. von Hoboken, *op.cit*.

[163] *Ibid.*

[164] For a more detailed analysis of a rights-respecting approach to the monitoring and removal of child abuse images, see Prostasia, Child protection and the limits of censorship, July 2018.

[165] For an analysis of the implications of *Eva Glawischnig-Piesczek v Facebook Ireland Limited,* C-18/18, see G. Smith, Notice and stay down order and impact on online platforms, *Bird &* Bird, October 2019. See also D. Keller, The CJEU's new filtering case, the Terrorist Content Regulation, and the Future of Filtering Mandates in the EU, The Center for Internet and Society Blog, 2 December 2019.

[166] For an example of what these transparency obligations might look like, inspiration could be taken from the proposed German law on media diversity rules for social media platforms, see Germany proposes Europe's first diversity rules for social media platforms. *op. cit.*

[167] M. MacCarthy, Transparency requirements for digital social media platforms: Recommendations for policy makers and industry, Working Paper, Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression , 12 February 2020.

[168] *C.f.* French Facebook Mission report, *op.cit.*

[169] The companies themselves tend to describe this as the 'prevalence' of content, i.e. the number of times it has been viewed.

[170] The Santa Clara Principles on Transparency and Accountability of Content Moderation Practices, 2017.

[171] See e.g. The European Commission and Member States consumer authorities ask social media companies to comply with European consumer rules, 17 March 2017.

[172] M. MacCarthy, *op.cit.*

[173] *C.f.* Sidestepping Rights, *op.cit.*

[174] See ARTICLE 19 policy on media diversity in the digital ecosystem, forthcoming.

[175] See an updated ARTICLE 19 policy on freedom of expression and elections, forthcoming.

[176] For a discussion of the impact of compliance targets, see Facebook, Charting a way forward, Online content regulation, February 2020.

[177] *Ibid.*

[178] See, e.g. V. Baines, On online harms and folk devils: Careful now, 24 June 2019.

[179] *C.f.* European Court of Human Rights, *Appleby and others v. The United Kingdom*, App. No. 44306/98, 6 May 2003.

[180] See, in particular. H. Tworek, R.Ó Fathaigh, L.Bruggeman & C.Tenove, Dispute resolution and content moderation: Fair, accountable, independent, transparent, and effective, Working Paper, Transatlantic Working Group on Content Moderation Online and Freedom of Expression, January 2020.

[181] See, e.g., H. Feld, The Case for the Digital Platform Act: Market structure and regulation of digital platforms, May 2019.

[182] For more details, see M. Masnick, Protocols, not platforms: A technological approach to Free speech, 21 August 2019; P.F. Docquir and M.L. Stasi, The decline of media diversity and how we can save it, 21 January 2020; M.L. Stasi, Ensuring pluralism in social media markets: Some suggestions, EUI Working Papers, RSCAS 2020/05, February 2020.