

Freedom of Expression in the Digital Public Sphere

Strategies for bridging information and accountability gaps in algorithmic content moderation

AUTHORS

Josh Cowls, Oxford Internet Institute, University of Oxford, UK

Philipp Darius, Centre for Digital Governance - Hertie School, Germany

Valentina Golunova, University of Maastricht, Netherlands

Sunimal Mendis, TILT - University of Tilburg, Netherlands

Erich Prem, Eutema Technology Management, Austria

Dominiquo Santistevan, University of Chicago, USA

Wayne Wei Wang, University of Hong Kong, Hong Kong SAR, China

This policy brief was formulated within the framework of the Research Sprint on *AI and Platform Governance* organized by the Alexander von Humboldt Institute for Internet and Society (HIIG) Berlin, Germany (August-October 2020). All authors contributed equally to the formulation of the policy brief.

Executive summary

A substantial portion of contemporary public discourse and social interaction is conducted over online social media platforms, such as Facebook, YouTube, Reddit, and TikTok. Accordingly, these platforms form a core component of the **digital public sphere** which, although subject to private ownership, constitute a digital **infrastructural resource** that is open to members of the public. As private entities, platforms can set their own rules for participation, in the form of terms of service, community standards, and other guidelines. The content moderation systems deployed by such platforms to ensure that content posted on the platform complies with these terms, conditions, and standards have the potential to influence and shape **public discourse** by mediating what members of the public are able to see, hear, and say online. Over time, these rules may have a norm-setting effect, shaping the conduct and expectations of users about what “acceptable” discourse looks like. Thus, the design and implementation of content moderation systems have a powerful impact on the **freedom of expression** of users and their **access to dialogic interaction** on the platform. With great power comes great responsibility: the increasing trend towards the adoption of algorithmic content moderation systems that have a questionable track record as regards their ability to safeguard freedom of expression gives rise to urgent concerns on the need to ensure that content moderation is regulated in a manner that safeguards and fosters robust public discourse in the online sphere.

This policy brief was developed with the contributions of an interdisciplinary team of experts from the fields of law, political science, sociology, and engineering. It aims to inform legislators and policymakers of the risks posed by the proliferation of algorithmic content moderation and the need for a more proactive regulatory approach by the states towards the governance of content moderation systems that are deployed by online platforms providing digital infrastructure for public discourse. It highlights an “information gap” that prevents regulators from evaluating the impact of content moderation on freedom of expression and an “accountability gap” that arises through the absence of effective redress mechanisms by which users are able to challenge violations of freedom of expression.

The policy brief concludes by proposing strategies for bridging these information and accountability gaps by means of:

Introducing enforceable statutory obligations which require platforms to:

- Provide information on the design, implementation, and impact of their content moderation systems through periodic reports and fundamental rights impact assessments (designed to address the information gap).
- Increase the capacity for individual redress through effective complaints and redress mechanisms (designed to address the accountability gap).

Establishing an Ombudsperson who is authorized to supervise the protection of freedom of expression on online platforms.

Facilitating informed, society-wide debate about how content moderation systems should be designed and optimised.

Platforms as digital infrastructure

A social media platform constitutes an online digital space, which enables and facilitates communication and interaction among a multiplicity of persons and/or entities by providing the necessary technical infrastructure and tools.¹ Accordingly, platforms give rise to a community of users who use this infrastructure to engage in public discourse by means of posting and sharing diverse types of content, thereby facilitating civic participation in addressing issues of public concern.² Political discussion and news media consumption, two substantial components of the social media experience, are particularly integral to civic participation³. Social media platforms' openness to the public, high number of users, and influence on contemporary social, political, and cultural discourse make it possible to describe them as a key component of the **digital public sphere**.⁴ Despite their importance for public discourse, platforms are owned and administered by private companies who exercise substantial discretion in determining the terms and conditions that govern user activity and behaviour on the platform. As platforms' user bases grow (to over 2.7 billion users in the case of Facebook), so do concerns over the influence of private decision-making on shaping public discourse. Thus, governments that support deliberative discussion, self-determination, and inclusive civic participation in the digital public sphere have a duty to facilitate and foster healthy and robust public discourse by ensuring greater fairness and transparency in platform governance.

Content moderation and freedom of expression

Content moderation in particular has considerable potential to shape and influence the nature and scope of the discourse taking place over a platform.⁵ Content moderation, as it concerns us here, is the process by which platforms shape community participation, prevent abuse, and ensure compliance with terms of service and community standards that they define.⁶ Platform owners and/or administrators are able to engage in content moderation by means of exercising their rights of ownership of the platform infrastructure and authority over its use⁷. It is the implementation of those terms, standards, and guidelines that are the single means by which rules are experienced by ordinary users.

Content moderation can come in many forms. On Facebook, a post can be reported by another user for violation of specific rules, an automated algorithm can spot potentially illegal content before it is shared, or in extreme cases, a user can be banned completely, making the platform virtually inaccessible to that user. Notwithstanding the variety of forms, content moderation has one clear impact: a private entity

¹ Definition of social media platforms developed by Valentina Golunova and Sunimal Mendis within the framework of the Research Sprint on 'AI and Platforms Governance'.

² On the notion of infrastructure see James Grimmelman, 'The Virtues of Moderation' (2015) 17 *Yale Journal of Law and Technology* 42, at p.52 citing Brett M. Frischmann, *Infrastructure: The Social Value of Shared Resources* (OUP 2012).

³ See Dhavan V. Shah et al, 'Information and Expression in a Digital Age: Modeling Internet Effects on Civic Participation' (2005) 32 *Communication Research* 531.

⁴ The term 'digital public sphere' is of relatively recent origin. It is envisioned as, "[...] a communicative sphere provided or supported by online or social media – from websites to social network sites, weblogs and micro-blogs – where participation is open and freely available to everybody who is interested, where matters of common concern can be discussed, and where proceedings are visible to all." Mike S. Schäfer, 'The Digital Public Sphere', in Gianpietro Mazzoleni et al.(eds.), *The International Encyclopedia of Political Communication* (Wiley Blackwell 2015) 322..

⁵ See Adam Kramer, Jamie Guillory and Jeffrey Hancock, 'Experimental Evidence of Massive-Scale Emotional Contagion Through Social Networks' (2014) 111 *Proceedings of the National Academy of Sciences of the USA* 8788.

⁶ See Grimmelman (n 2) at p. 47 and definition of "Content Moderation" developed by Phillip Darius and Wei Yan within the framework of the Research Sprint on "AI and Platform Governance"..

⁷ Although it is noted that in some contexts, this freedom is theoretically limited and must not infringe on the users' individual rights – See Engle E, 'Third Party Effect of Fundamental Rights (Drittwirkung)' (2009) 5 *Hanse L Rev* 165.

makes a decision concerning an individual’s ability to access and to participate in the public discourse taking place over the platform and interacting with a community of users.

Content moderation therefore has significant implications for the protection of freedom of expression in the digital public sphere. While pursuing a legitimate objective of preventing abuse by taking down illicit or harmful content and ensuring compliance with the platforms’ terms of service and community standards, **inaccuracies and inefficiencies in the design and implementation** of the content moderation system can negatively impact freedom of expression *inter alia* in the following ways:

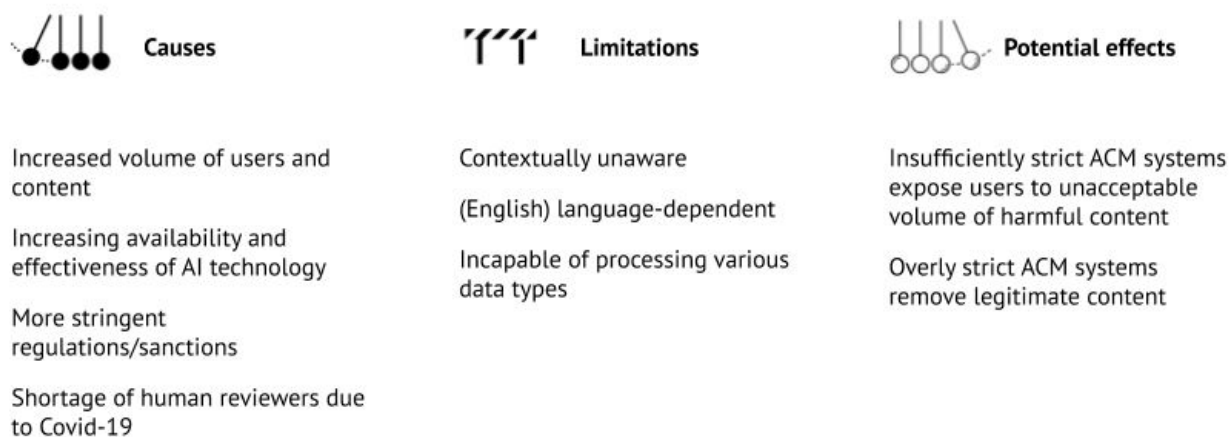
- Wrongful removal/flagging of legitimate content that impedes free-speech.
- Wrongful banning of users/disabling of accounts that prevents individuals from accessing and participating in the public discourse taking place on the platform.

What is even more troubling is that in certain instances, content that is *not* deemed illicit or harmful under prevailing law can be removed and user accounts disabled on grounds of non-compliance with the platform’s terms of service or community standards.

As exemplified below, the recent trend towards the deployment of algorithmic content moderation (ACM) systems exacerbates the negative impact of content moderation on freedom of expression.

The proliferation of algorithmic content moderation systems

THE PROLIFERATION OF ACM SYSTEMS



Causes

Many online networks have turned to computational solutions that help manage content contributed by users. These systems use algorithms for the classification of content to support platform moderation. The algorithms may classify content as illegal or potentially harmful, or they may be used for the identification of specified content, e.g. copyright-protected material. The classification of content can then be used for automated deletion or blocking, or as a support tool for human moderators (i.e. pre-moderation) who make the final decision regarding the illegality or harmfulness of the content.

The automated classification of content is a huge challenge. Even the recognition of predefined material (e.g. online music) is a complex task due to the large amounts of data, the necessary availability of large databases, issues of interoperability, and the requirement to deal with modified content, e.g. truncated or noisy files. The classification of illegal online content or of potentially harmful content also poses considerable technical challenges as it targets a more semantic level of the content, i.e. its meaning. These systems use different algorithms, including so-called artificially intelligent techniques. They range from rather simple schemes that focus on the identification of key words to more elaborate systems that use deep learning methods to identify postings with potentially illegal content based on large sets of examples previously classified by human experts. Despite efforts to make these systems more perceptive of nuance, most of them remain context-blind and therefore produce highly inaccurate results.⁸

The current state-of-the-technology is far from the capabilities and recognition rate of humans in many domains. Existing methods can only achieve partial recognition. This is unsurprising given that the inter-expert agreement rate for classifying online content varies significantly. Most algorithmic methods only work on single contributions, such as a single post, and cannot properly deal with the overall context of postings, such as a chat history, its news context, timely aspects or cultural connotations. Typically, classification algorithms therefore can only deliver a rough classification quality with relatively large numbers of false positives and false negatives.

The increased use of automated content moderation can be attributed to several causes. This includes the surge in the volume of content that is posted and shared over online platforms due to increasing numbers of users and user activity, and as we have noted, the increasing availability of new artificial intelligence technology that is perceived to be more speedy, efficient, and cost-effective than human moderators.

However, the surge in the use of ACM can also be, in part, attributed to more stringent regulatory frameworks that impose **greater pressure on platforms to monitor and filter content**. Traditionally, platforms were either shielded from liability for third-party content (e.g. Section 230 of the US Communication Decency Act⁹) or held liable in accordance with a 'negligence-based' approach, whereby liability was imputed only where they failed to take down illegal or harmful content once they were made aware of its existence on the platform (e.g. Arts. 14 & 15 of the EU e-Commerce Directive¹⁰ and Art. 1195 of the Chinese Civil Code¹¹).

Nevertheless, at present, there is a **global shift towards increasing the degree of liability** imposed on platforms for hosting illegal/harmful content as a means of compelling platforms to take a more proactive role in combating the surge of illegal content online. This shift can be described as being aimed towards coercing or co-opting platform owners into regulating content shared on their platforms by third parties on behalf of the state.¹² At the extreme end of the spectrum of this shift are laws that impose strict liability regimes. For instance, under Article 17 of the EU C-DSM Directive¹³, online content-sharing service providers are held directly liable for copyright-infringing content posted on their

⁸ See Tarleton Gillespie, 'Content Moderation, AI, and the Question of Scale' (2020) 7 *Big Data & Society* 1.

⁹ Communications Act of 1934 at 47 U.S.C. § 230 (promulgated under the Telecommunications Act of 1996).

¹⁰ 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 [2000] OJ L 178/1 (hereinafter "EU e-Commerce Directive").

¹¹ The Civil Code of the People's Republic of China, as the first law named after the Code, was promulgated on May 28, 2020 and will come into force on January 1, 2021 (hereinafter "the Chinese Civil Code", CCC).

¹² See Jack M Balkin, 'Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation' (2018) 51 *UC Davis Law Review* 1149, at p.1153.

¹³ Directive 2019/790/EU 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 [2019] OJ L 130/92 (hereinafter "EU C-DSM Directive").

platforms by third parties. A strict liability regime is also envisioned in the Proposed Regulation on Terrorist Content.¹⁴ The need to avoid the risk of incurring liability can serve as a strong incentive for platforms to deploy ACM systems in order to ensure proactive detection and removal of illegal/harmful content.

The deployment of ACM is also induced due to **the lack of the ‘Good Samaritan’ protection**. For example, under Section 230 of the US Communications Decency Act, platforms cannot be held liable for content which they take down voluntarily provided they do so in good faith. Importantly, platforms are equally protected against liability for illegal content which they had undertaken to tackle, but failed to detect or remove.¹⁵ However, the ‘Good Samaritan’ protection is absent in the legislative framework of other jurisdictions, including the EU¹⁶ and China.¹⁷ Voluntary moderation efforts can be regarded as an indication that platforms do not store content in a mere passive and automatic manner and lead to the waiver of the liability exemption. As the cost of a single mistake can be too high, platforms implement ACM systems, which are believed to surpass relevant human capabilities. For example, the Good Samaritan gap between China and the US triggered America’s National Security review of TikTok. – see TikTok Case Study.

Platforms are also incentivised to increase the efficiency of their content-filtering activities through the imposition of **monetary penalties**. For instance, the German NetzDG Law¹⁸ foresees significant fines of up to 50 million euros. In a similar vein, Article 68 of the Chinese Cybersecurity Law¹⁹ envisages economic fines, suspension of business, and monetary penalties for persons in charge. The introduction of such stringent regimes induce platforms to resort to ACM as a more effective means of compliance.

In addition to more stringent policies, the increased use of ACM can also be attributed to other factors. Apart from a general rise in user-generated online content, one of the many effects of the **global COVID-19 pandemic** was the inability of employees to work from their office. For large platforms, such as Facebook or YouTube, that perform *ex post* content moderation in-house, this meant that tasks that would normally be done by human moderators were performed by automated algorithms in the interim. These unusual circumstances provide a unique view into the effects of increased ACM. Despite advancements in the capacity and sophistication of ACM technology, our research leads us to conclude that increased use of ACM without substantial human oversight results in higher levels of wrongful removal and disabling of legitimate content – see YouTube Case Study.

¹⁴ Proposal for a regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online (COM(2018) 640, 12 September 2018).

¹⁵ See Aleksandra Kuczerawy, ‘The EU Commission on Voluntary Monitoring: Good Samaritan 2.0 or Good Samaritan 0.5?’ (CITIP blog, 24 April 2018) <<https://www.law.kuleuven.be/citip/blog/the-eu-commission-on-voluntary-monitoring-good-samaritan-2-0-or-good-samaritan-0-5/>> accessed 2 October 2020.

¹⁶ For instance, Articles 14 and 15 of 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) provide a “safe harbour” for intermediaries who ‘host’ content (as opposed to merely caching or transmitting content) only on the basis that they comply with notice-and-take-down obligations.

¹⁷ Although the Articles 1194-1197 of the Chinese Civil Code (CCC), on the basis of Article 36 of the previous Tort Law, integrate the “counter-notice” provisions of the Articles 42 and 43 of the Electronic Commerce Law, the Article 1195 (CCC) still stipulates that after receiving the notice, the ISPs shall promptly transmit the notice to the relevant network users, and take necessary measures based on the preliminary evidence of infringement and the type of service.

¹⁸ German Network Enforcement Act [Netzwerkdurchsetzungsgesetz] BGBl. I S. 3352 (2017) updated as BGBl. I S. 1328, 1360 (2020).

¹⁹ The Cyber Security Law of the People’s Republic of China was enacted on November 7, 2016 and implemented on June 1, 2017. See the English translation of the Chinese Cybersecurity Law at <https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-cybersecurity-law-peoples-republic-china/>.

Effects

The increased reliance on ACM has created a wellspring of effects for the social media ecosystem. The core challenge of creating an ACM system is to strike an appropriate balance between protecting users against exposure to objectionable (i.e. illegal or harmful) content and ensuring that users' freedom of expression is adequately safeguarded. An insufficiently effective system might allow an unacceptable level of objectionable content to remain on the platform, potentially exposing users to considerable harm, offence, or risks to safety. On the flipside, however, an overzealous ACM system may unfairly remove legitimate content, creating a chilling effect with respect to what is seen as acceptable. Here we examine the latter concern and identify three primary limitations of algorithmic moderation technology that negatively impact on safeguarding users' freedom of expression.

First, **ACM systems are context-blind**. While capable of detecting certain words and expressions, they interpret them as mere strings of data rather than components of human lexicon. For example, an ACM tool targeting obscenity online cannot ascertain whether an individual used strong vocabulary to incite hatred or to raise an essential matter of public interest. The ignorance of nuance can threaten forms of online expression that contribute to important public debates.²⁰

Second, **ACM is significantly language-dependent**. Most systems powered by natural language processing are only trained to process texts in English and a few other frequently used languages. This makes their accuracy rate much lower than when they are deployed for analysing texts written in more uncommon languages.²¹ Furthermore, most platforms are unable to review automatically flagged materials written in less common languages as they lack moderators fluent in them.²² As a result, the roll-out of ACM can impair the freedom of expression of certain language speakers, especially those belonging to linguistic minorities for which only relatively small training-datasets are available.

Third, **ACM is incapable of processing certain types of files**. Most ACM tools are only able to process conventional types of files, such as text images and videos. However, they are still incapable of ensuring adequate analysis of other viral types of content, including memes and GIFs.²³ Due to the low accuracy of ACM, such materials may either remain undetected or be subject to over-removal. The limited capabilities of ACM tools can therefore give rise to an asymmetry among various forms of online expression.

Closing the information gap

People can reasonably disagree about how to balance freedom of speech with other interests and how these different priorities should be weighed. All can agree, however, that this is not a decision for platforms to take by themselves. What is needed instead is informed, inclusive debate among all stakeholders, backed up by responsive policymaking, to ensure that all perspectives are heard.

Platforms have begun to take such steps, at least internally. Facebook has a large public policy team which engages with experts and other key stakeholders around the world in order to develop its community standards and assess how well they work in practice. It also analyses user appeals to get a

²⁰ Otto Preminger-Institut v Austria [1994] App. No. 13470/87, para 49; Marina v Romania [2020] App. No. 50469/14, para 74; Gündüz v Turkey [2003] App. No. 25071/97, para 37;

²¹ See Natasha Duarte and others, 'Mixed Messages? The Limits of Automated Social Media Content Analysis' in Proceedings of the 1st Conference on Fairness, Accountability and Transparency (PMLR 2018).

²² See Olivia Solon, 'Facebook's Failure in Myanmar is the Work of a Blundering Toddler' (the Guardian, 16 August 2018) <<http://www.theguardian.com/technology/2018/aug/16/facebook-myanmar-failure-blundering-toddler>> accessed 25 October 2020.

²³ See Bertie Vidgen and others, 'Challenges and Frontiers in Abusive Content Detection' in Proceedings of the Third Workshop on Abusive Language Online (Association for Computational Linguistics 2019).

sense of the types of content, which users are most likely to believe should remain on the platform. It has also established an independent Oversight Board which will adjudicate on individual cases in order to develop principles that can, in turn, be used to reshape its community standards, and even hear appeals from users of other platforms.²⁴

Such steps are welcome, but they remain in large part a closed loop, with little ability for ordinary users to engage in the process. And although Facebook has for several years been releasing high-level data about its moderation of various kinds of content in the form of Transparency Reports (see Facebook Case Study), this information lacks sufficient granularity for outside researchers to understand how and why content moderation decisions are made. On the individual level, the amount of information provided to users to enable them to understand why a particular piece of content has been removed varies considerably among different platforms (see table below).

KEY FEATURES OF APPEALS PROCEDURES

Summary of key features of appeals procedures of four large social media platforms and the online discussion forum of the Austrian newspaper, Der Standard (see Der Standard Case Study)

	User gets notified about content removal decisions?	Appeal mechanism visible?	Appeal mechanism visible?	Appeal mechanism visible?
Facebook	Yes	On site	Yes	Yes
Reddit	Sometimes, but not enforced	Case by case with moderator	Yes	Yes
TikTok	Yes, notification in the application	Integrated in the application	No	No
YouTube	Yes	In YouTube Studio	Yes	Yes
Der Standard	Indirectly using a counter	Yes, email address that needs to be found on the website	No report, but some figures published	No

This information gap – between what platforms know and what individual users, regulators, and other stakeholders and interested parties, such as scientific researchers, know – urgently needs to be closed. Only when all stakeholders have access to sufficient data can informed and inclusive debate take place about how content moderation systems ought to work and the capacity of ACM systems’ fitness for safeguarding freedom of expression be assessed.

²⁴ At the same time it is worth mentioning that Facebook aims for a single set of globally applicable standards.

Closing the accountability gap

In addition to the information gap we have also identified an accountability gap that arises from the absence of effective and transparent complaints and redress mechanisms which enable users to challenge wrongful removal of legitimate content and disabling access.

Not all legislative instruments targeted at regulating platforms explicitly require platforms to establish effective complaints and redress mechanisms. On the other hand, those that do impose such an obligation do not incorporate adequate measures to ensure the effectiveness of the complaints and redress mechanism. Furthermore, the legal enforceability of these obligations are unclear.

The EU C-DSM requires online content-sharing service providers to put in place “an effective and expeditious complaint and redress mechanism”. However, the enforceability of this obligation is uncertain, as no liability is apportioned for non-compliance. In the US, s. 512 (g) of the Digital Millennium Copyright Act (DMCA)²⁵ provides for a system of “counter-notice” by means of which a user is able to challenge the removal or disabling of content. However, the burden of establishing the lawfulness of the content is left squarely on the shoulders of the user, something that all users may not have the resources or expertise to perform. On the other hand, submitting a counter-notice can itself expose the user to an infringement law-suit; a risk many users will be unwilling to take.²⁶

An effective complaints and redress mechanism would constitute a powerful tool for safeguarding freedom of expression on online platforms, as it would afford users a means by which to challenge wrongful censorship of free-speech without incurring the costs and risks associated with court procedures. Thus, closing the accountability gap requires the introduction of positive and enforceable obligations on platforms to provide such mechanisms and adequate regulatory supervision to ensure their efficacy. To take China’s online content moderation as an example, some ministry-level regulatory documents that are legally binding require platforms to “practice social responsibility and industrial self-discipline”, as well as “improving the mechanisms for screening malicious reports and dealing with complaints in a timely and fair manner”,²⁷ but it is unclear whether the obligation should be applicable to ACM systems. China also called for the “net norms” as the self-governance instruments of platforms in the main ministry-level internet-regulating documents, which encouraged most platforms to adopt complaints and redress mechanisms.²⁸

Proposals




We propose the following measures with the objectives of bridging the information and accountability gaps exemplified above and of safeguarding freedom of expression on online platforms that are open to public participation and interaction.

²⁵ Digital Millennium Copyright Act (USA) 17 U.S. Code § 512 (hereinafter “DMCA”).

²⁶ DMCA § 512(g)(2)(C).

²⁷ Arts 14 and 15 of the Internet User Public Account Information Services Management Provisions issued by Cyberspace Administration of China (Promulgated in 2017). See the English translation of the document at <https://www.chinalawtranslate.com/en/public-accounts/>.



²⁸ For instance, in the WeChat Public Account Platform, when the relevant content is suspected of violating the law and is moderated (often deleted), a “complaint initiation” link will be attached to the decision notice to the users concerned.

	ACCOUNTABILITY GAP	INFORMATION GAP
 Appropriate court	Legal protection <hr/> Adjudicate appeals against Ombudsperson's decisions on EU or national level	
 Ombudsperson	Powers of review <hr/> Review decisions delivered by platform concerning violations of FoE	Supervise, evaluate, and recommend
 Platforms	Enforceable statutory obligation <hr/> Establish and maintain effective complaints and redress mechanisms	Periodic reports on number of complaints received and reinstatements carried out Fundamental rights impact assessment to assess impact of ACM systems on freedom of expression

The goal of these proposals is to strengthen the character of publicly accessible online platforms as digital spaces for robust and healthy public discourse by:

- Urging regulators to take a more proactive role in supervising and regulating content moderation by privately owned platforms, especially in relation to the protection of freedom of expression (this is contrasted with the existing approach, whereby states favour self-regulation by platforms with minimal public supervision).
- Requiring greater openness and transparency from platforms with regard to their content moderation processes, particularly where ACM systems are deployed.
- Enhancing the accountability of the content moderation process by empowering users to challenge wrongful take-down of content which may violate users' freedom of expression.

Our proposals consist of two principal layers:

-  Introduction of **enforceable statutory obligations** requiring platforms to establish effective complaints and redress mechanisms (accountability gap), periodic reporting requirements, and fundamental rights impact assessments (FRIAs) (information gap).
-  Establishment of an **Ombudsperson**,²⁹ a public authority vested with powers to supervise and assess the safeguarding of freedom of expression by platforms and to review decisions delivered by the platform pursuant to the complaints and redress process, for violations of freedom of expression.

²⁹ The office of the Ombudsperson is contrasted with existing private oversight mechanisms that have been established by platforms, the best example of which is the Facebook Oversight Board. We believe that entrusting the supervision of content moderation systems deployed by platforms to a public authority will lead to greater transparency and accountability.

Proposals for closing the accountability gap

Enforceable statutory obligations

We propose the introduction of a statutory obligation to establish and maintain **effective complaints & redress mechanisms**. This obligation would apply to all online platforms that host content provided by members of the public. Failure to comply with the obligation would expose the platform to state-imposed sanctions and/or administrative penalties imposed by the Ombudsperson.

The statutory obligation would require the platform to ensure that their appeals and redress mechanism incorporates *inter alia* the following features:

- ✓ Duty to notify users of removal/disablement of content posted by them (including “shadow banning³⁰”).
- ✓ Consideration and determination of complaints to be carried out by human reviewers.
- ✓ Reviewers to be provided with adequate training and resources to consider and determine complaints.
- ✓ Right of users and other interested parties to be heard in determining the complaint.
- ✓ Complaints to be scrutinized and decision delivered and (if applicable) redress provided within a defined time-frame that is determined in a manner that balances the interests of platform owners and users.
- ✓ Duty to give reasons for decision.

While acknowledging that compliance with this obligation would impose additional financial and administrative burdens that could be onerous for small and medium-sized platforms, we believe that they are necessary and proportionate for ensuring the protection of freedom of expression. Given the character of these platforms as digital public infrastructures, it is reasonable to require them to put in place minimal safeguards to ensure that content moderation is carried out in a manner that safeguards due process and user rights and interests.

Ombudsperson

The Ombudsperson will be authorised to **review decisions delivered by platforms** pursuant to the complaints mechanism, where the aggrieved party is of the view that the decision violates their freedom of expression. In such a case, the Ombudsperson will review the decision for possible violations of freedom of expression and will deliver a ruling that is final and binding on all parties. The Ombudsperson will be authorised to order restitution (e.g. content to be restored, account to be unblocked) *and*, where necessary, the payment of damages/compensation to the aggrieved party.

Furthermore, where the human reviewers dealing with complaints are uncertain about determining the freedom of expression issues that arise through a complaint, they will be able to consult with the Ombudsperson who will provide advice and guidance on determining those issues.

A platform’s failure to comply with a ruling delivered by the Ombudsperson will expose it to administrative penalties and/or legal sanctions. Any party who is dissatisfied with a ruling issued by the Ombudsperson will have the possibility of appealing the ruling to an appropriate national or regional court.

It is anticipated that this measure would ensure that determinations on potential violations of freedom of expression are not exclusively left to private entities, but are subject to oversight by a public authority who will be qualified to engage in a comprehensive analysis of the legal issues involved. Moreover, it

³⁰ Shadow banning refers to the blocking of a user or content posted by a user, from the platform in a manner that does not make it evident to the user that he has been banned.

would resolve the information asymmetry problem which refers to the absence of adequate knowledge and skills on the part of human reviewers employed by platforms to correctly assess the legal issues relating to the violation of freedom of expression.

Proposals for closing the information gap

In addition to the general obligation proposed above we propose the introduction of two additional enforceable statutory obligations that would apply to platforms that are considered to have a particularly significant impact on public discourse, e.g. by means of catering to a large community of users, attracting a large public audience, or hosting a large volume of content.

The determination of the exact criteria based on which these obligations will become applicable should be left to policy-makers and regulators. However, existing legal provisions differentiate levels of liability based on factors, such as number of monthly unique visitors³¹, annual turnover of the platform³², and so on. It is also possible that not-for-profit educational and scientific platforms and open-source software developing platforms will be exempted from these specific obligations taking into account the financial and administrative burdens that they entail. The objective of this limitation is to exempt smaller and less-financially profitable platforms from the financial and administrative commitment involved in complying with the statutory obligations proposed below. On the other hand, it is reasonable to require large platforms that profit from the public interaction and discourse taking place on the online infrastructure provided by them to take reasonable and proportionate measures to ensure that the content moderation systems deployed by them safeguard user interests and freedom of expression to an adequate extent.

Enforceable statutory obligations

First, we propose an enforceable statutory obligation to provide **periodic (annual/biannual) reports** to the Ombudsperson on the number of complaints received by platform and number of reinstatements carried out. This proposal is inspired by Section 2 of the NetzDG Law which introduces a similar half-yearly reporting obligation. It is envisioned that this reporting obligation would provide greater transparency to the complaints and redress process and enable regulators and members of the public to monitor and evaluate the accuracy of the content moderation systems and the efficacy of the complaints and redress mechanism.

The report should *inter alia*:

- ✓ Categorise complaints according to the type of action that was challenged (e.g. removal, disablement of access), reason for action-taken (e.g. hate-speech, copyright infringement) and whether complaint was made by an individual user or an organization.
- ✓ Categorise reasons for reinstatement (e.g. human error, error on the part of the AI system, decision overturned by Ombudsperson on appeal).
- ✓ Indicate the time-span within which removal/disablement of content was made, complaint was received, decision provided, and reinstatement carried out.
- ✓ Indicate the number of decisions that were appealed to the Ombudsperson.

Second, we propose an enforceable statutory obligation to **provide an annual FRIA** to assess the impact of the deployment of ACM on freedom of expression. The aim of the FRIA is to enable regulators and

³¹ EU C-DSM Art. 6

³² Ibid.

members of the public to evaluate and supervise the measures taken by a platform for safeguarding freedom of expression. It is inspired by the Data Protection Impact Assessment that was introduced under Article 35 of the EU GDPR³³. Such an impact assessment would be especially relevant for assessing the impact of new technologies, such as automated processing systems, to fundamental rights and freedoms.³⁴

The FRIA should *inter alia* incorporate information on the following aspects:

- ✓ How do the community standards/ terms and conditions of the platform safeguard freedom of expression?
- ✓ Description of ACM systems used, the purpose of the use, and the mode of implementation.
- ✓ Assessment of the necessity and proportionality of employing ACM systems in relation to the stated purpose.
- ✓ Identification of the risks of AMC systems to freedom of expression of users and measures employed to mitigate risks.
- ✓ Description of the appeals procedure and remedial measures provided to users.
- ✓ Description of training and resources provided to human reviewers (dealing with complaints) to carry out a comprehensive legal analysis of the freedom of expression issues involved.

Ombudsperson

Regarding the second layer of our proposals on closing the information gap, the Ombudsperson will be authorized to:

- Receive and assess the periodic reports and FRIAs submitted by platforms and issue recommendations for improving content moderation systems and complaints and redress procedures with the objective of better protecting freedom of expression.
- Impose administrative penalties for failure to comply with statutory obligations or to implement recommendations.
- Maintain a website that is available to the public where summaries of periodic reports and FRIAs provided by platforms are published as well as rulings or summaries of rulings delivered by the Ombudsperson in the exercise of his powers of review, which are redacted as required by law (e.g. to protect personal information or commercially-sensitive information).
- Provide guidance and advice to platforms on their statutory duties and in ensuring that their content moderation systems comply with statutory requirements.

³³ Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, OJ 2016 L 119/1(hereinafter EU GDPR)

³⁴ Our proposal has been inspired by Janssen's proposal for the introduction of an FRIA for automated decision making with personal data. See Heleen L Janssen, 'An Approach for a Fundamental Rights Impact Assessment to Automated Decision-Making' (2020) 10 International Data Privacy Law 76.

A platform's failure to comply with statutory obligations or to implement a recommendation delivered by the Ombudsperson will expose it to legal sanctions and/or administrative penalties imposed by the Ombudsperson. Any party who is dissatisfied with a ruling issued by the Ombudsperson will have the possibility of appealing the ruling to an appropriate national or regional court.

It is anticipated that close supervision of the fitness of the content moderation systems (particularly ACM systems) by the Ombudsperson and periodic evaluation thereof will incentivize and compel platforms to ensure that those systems are designed and calibrated to avoid violations to freedom of expression through the removal or disabling of legitimate content.

The publication of summaries of the periodic reports and FRIAs as well as rulings (or summaries thereof) on a publicly available website by the Ombudsperson is expected to ensure the availability of information on the efficacy and accuracy of content moderation to stakeholders, interested parties (e.g. scientific researchers), and the wider public. It is hoped that this would contribute towards facilitating informed, society-wide debate on user freedoms and the calibration of content moderation systems to safeguard and foster freedom of expression in the digital public sphere.

Case Studies

YouTube Case Study

YouTube, the largest video hosting platform on the internet, has been the target of recent criticism by governments and users concerning their use of algorithms. Politicians have accused YouTube of political bias, algorithmic reinforcement of extremist views by recommendation systems, and the circulation of misinformation, while content creators, some of whom rely on revenue from videos as a main source of income, have strongly censured the platform for their lack of accountability on decisions of removal. The majority of these decisions are made by automated algorithms and in the wake of the pandemic and the absence of human moderators, YouTube relied more heavily on these algorithms.

Through numbers made available in YouTube's transparency reports and conversation with policy representatives, we were able to investigate the effects of increased ACM more closely. As we see in the table below, over 11 million videos were removed in quarter 2 of 2020, almost double the number removed in the first quarter. From the drastic increase in removals and the disproportionate increase in reinstatements, we inferred that the **algorithms are simply not able to perform as well as human moderators**. This hypothesis was confirmed by a YouTube representative. Despite advances in the technological capability of automated algorithms, we conclude that more automated moderation without a human in the loop is not the solution.

Removal, appeals, and reinstatement per quarter	Total number of videos removed	Appeals share	Reinstatement share
Q4 2019	5887021	1,81%	19,6%
Q1 2020	6111008	2,72%	24,7%
Q2 2020	11401696	2,85%	49,4%

Table: Removals, appeals, and reinstatements before and during the coronavirus pandemic (Data: Google Transparency Report)

Facebook Case Study

Content moderation on Facebook: what do we know?

Calls for Facebook to standardise and improve its content moderation have naturally been accompanied by demands to know whether and to what extent Facebook is actually doing so. Facebook's community standards transparency reports, which provide data stretching as far back as late 2017, provide basic information about the scope and type of violating content that it "actioned" – i.e., removed – from its main platform as well as from its popular subsidiary Instagram. More specifically, the data provided is as follows:

Prevalence – showing in absolute terms the volume of problematic content in a variety of categories; however, full longitudinal data is only available for the "Adult Nudity and Sexual Activity" and "Violent and Graphic Content" categories. For other categories, Facebook claims that either the volume of violating content seen by users is either too small to create a robust sample, or simply that data cannot be estimated because "our prevalence measurement is slowly expanding to cover more languages and regions". In still other cases, the data is reported for a single quarter, e.g. "we estimate that fake accounts represented approximately 5% of our worldwide monthly active users (MAU) on Facebook during Q2 2020."

Content actioned – showing the volume of content that was “actioned” i.e. removed by Facebook, for each quarter.

Proactive rate – showing the percentage of content that was found and flagged by Facebook before being reported by users for each quarter. In almost all cases this “finding and flagging” is done automatically using rules-based and/or AI detection systems.

Appealed content – showing in absolute terms how much content that was removed was then appealed by users, for each quarter.

Restored content – showing in absolute terms how much content was restored, either without an appeal (i.e. following internal review) or after a successful appeal.

Below are graphs showing the content actioned, proactive rate, appealed content, and restored content for hate speech, an especially challenging category of content with respect to enforcement and the balancing of users’ rights and expectations. Unfortunately, the data provided in the transparency reports is not offered in an open or machine-readable format, so the graphs below are merely screenshots from the Transparency Report.



There are undoubtedly several things we can learn from the data that Facebook releases publicly. Basic longitudinal data highlights several trends with respect to the overall prevalence of content (at least for some categories) and to the responses Facebook takes. The data allows us to at least informedly

speculate about what underlies these trends, such as the impact of the COVID-19 pandemic and the rise of automated flagging technology. Yet there are several questions that an interested user would be unable to answer without the points of contact that the authors were able to draw upon. First, there is a lot of missing data, particularly with respect to prevalence, with only two of the ten content categories offering full longitudinal data. Second, the format that the data is presented in prevents more granular analysis without laborious manual recreation of the data in a manipulable format. Third and most importantly, the reports say little about *how and why* decisions get made in the way that they do. As a recent piece by the EFF argues,³⁵ “true transparency must provide context”, and there is little context in the transparency reports regarding the basis on which individual decisions are made. Nor is information available about the process of content moderation itself, the use of automated technology, besides that which can be inferred from the above (let alone how such technology works), or the background and location of content moderators. Recent changes made to Facebook’s Community Standards³⁶ can be found on its website, but again, *how and why* these changes were made is not made clear – and nor is the relationship between the development and iteration of standards and how they are enforced apparent.

Finally, the impact of the COVID-19 pandemic appears to have been stark. Facebook effectively suspended its appeals process as mass lockdowns went into effect given the limited resource available, explaining the drastic drop in appeals, though it appears to have restored much more content outside of the appeal process, which more than makes up for the previous quarter’s total. Of course, it is impossible to assess the accuracy of the decision to restore content in this way.

TikTok Case Study

TikTok/Douyin per se is a split product. Since 2019, ByteDance has been separating TikTok International from Douyin (TikTok’s Chinese version),³⁷ but US TikTok has still been at the core of the debate on trans-border data flow and ACM in the “Splinternet”. At the beginning, when ByteDance was a start-up in China, Douyin was successful in shaping its business model with its core advantaged algorithms, and this model originated from that of ByteDance’s Toutiao, a news app, which controls algorithmic news feed based on user behavior as training data to algorithms.³⁸

TikTok/Douyin was originally algorithmically moderating creative contents to personalize the recommendations on a large scale. The data were thus the fuel to train algorithms. Subsequently, the Good Samaritan gap between China and the US triggered America’s review of TikTok.³⁹ One reason may be that in the beginning of the entry to the US market, the Good Samaritan gap was constituted between TikTok’s and Douyin’s ACM systems and the latter was shaped by the more stringent ACM built upon Chinese legislative packages relatively lacking in the articles encouraging voluntary moderation efforts.

³⁵ See Svea Windwehr and Jillian C. York, ‘Thank You For Your Transparency Report, Here’s Everything That’s Missing’ (Electronic Frontier Foundation, 13 October 2020)
<<https://www.eff.org/deeplinks/2020/10/thank-you-your-transparency-report-heres-everything-thats-missing>> accessed 30 October 2020.

³⁶ See Facebook Community Standards at <https://www.facebook.com/communitystandards/>.

³⁷ See The Technology Policy Institute, “‘TikTok Public Policy’s Michael Beckerman’ (Two Think Minimum)’ (2020)
<<https://techpolicyinstitute.org/2020/08/05/tiktok-public-policys-michael-beckerman-two-think-minimum/>> accessed 6 August 2020.

³⁸ See Ben Thompson, ‘The TikTok War’ (Stratechery by Ben Thompson, 14 July 2020)
<<https://stratechery.com/2020/the-tiktok-war/>> accessed 15 July 2020.

³⁹ See The White House, ‘Executive Order on Addressing the Threat Posed by TikTok’ (The White House, 2020)
<<https://www.whitehouse.gov/presidential-actions/executive-order-addressing-threat-posed-tiktok/>> accessed 2 October 2020.

In response to the doubts and questions, since 2019, TikTok has been building an internal “Jurisdictional Wall” to separate data, codes, and algorithms between jurisdictions.⁴⁰ While strengthening data protection, TikTok’s data separation/localization policy constitutes a self-governance model on ACM – limiting overseas access to the domestic training data for moderating algorithms. And TikTok’s ACM reform has also released its transparency reports and induced the Transparency & Accountability Center (TAC), TikTok Content Advisory Council,⁴¹ and the “Zhong Tai”.⁴² Due to the global pandemic, the details of the TAC were less revealed, but the members in the Content Advisory Council have represented the diversity and inclusiveness of different stakeholders in the US.

Der Standard Case Study

ACM is not limited to large online platforms. The Austrian newspaper Der Standard (paid circulation app. 56000, online visitors app. 35 million p.m.) runs a moderated online forum. Registered users comment on news articles and the moderators aim at an engaged and constructive discourse. They use two algorithmic support tools: the pre-moderation classification system *Foromat* trained to classify user comments in several categories, including hate speech, discrimination etc. The second system, the *de-escalation bot*, aims to identify postings that have a positive influence on the discussion.

The AI approach for classification in seven categories was developed and published by the Austrian Research Institute of Artificial Intelligence⁴³.

Der Standard uses algorithms mostly to deal with the large number of online content while maintaining quality of discourse. For the most part, the systems help human moderators to decide on posts that should not be published or not highlighted. However, the system is also capable of automatically removing content, e.g. duplicated entries. Format has been in use for more than 15 years.

Typically, moderators will check the posts with the help of the algorithms before they go online. In case of deletion, users can check a counter that informs them about removed content. However, user posts can also be deleted without being inappropriate, e.g. when they replied to an unsuited post and the counter cannot currently distinguish such cases.

Users can appeal the removal of their post by sending an email to a somewhat hidden address. The cases are then reviewed and the users receive a feedback email or their post is reinstated right away. Der Standard does not yet publish regular reports about deletions, complaints, or reinstatations. During the COVID-19 crisis, it became impossible for the moderators to deal with appeals due to the increasing number of posts while staff was on short labour.

⁴⁰ The Technology Policy Institute (n 38).

⁴¹ See Vanessa Pappas, ‘Introducing the TikTok Content Advisory Council’ (Newsroom | TikTok, 16 August 2019) <<https://newsroom.tiktok.com/en-us/introducing-the-tiktok-content-advisory-council>> accessed 2 October 2020.

⁴² “Zhong Tai”, or “中台” in Chinese, is a technical and business jargon that roughly means anything sandwiched between a tech company’s front-end and back-end. See Chen Du, ‘ByteDance Cuts Domestic Engineers’ Data Access to TikTok, Other Overseas Products’ (PingWest, 2020) <<https://en.pingwest.com/a/6875>> accessed 14 June 2020.

⁴³ See Dietmar Schabus and others, ‘One Million Posts: A Data Set of German Online Discussions’ in Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (Association for Computing Machinery 2017).

References

Articles and Research Papers

- Jack M Balkin, 'Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation' (2018) 51 UC Davis Law Review 1149, at p.1153.
- Natasha Duarte and others, 'Mixed Messages? The Limits of Automated Social Media Content Analysis' in Proceedings of the 1st Conference on Fairness, Accountability and Transparency (PMLR 2018).
- Tarleton Gillespie, 'Content Moderation, AI, and the Question of Scale' (2020) 7 Big Data & Society 1.
- James Grimmelman, 'The Virtues of Moderation' (2015) 17 Yale Journal of Law and Technology 42, at p.52 *citing* Brett M. Frischmann, *Infrastructure: The Social Value of Shared Resources* (OUP 201)
- Dhavan V. Shah et al, 'Information and Expression in a Digital Age: Modeling Internet Effects on Civic Participation' (2005) 32 Communication Research 531.
- Engle E, 'Third Party Effect of Fundamental Rights (Drittwirkung)' (2009) 5 Hanse L Rev 165.
- Adam Kramer, Jamie Guillory and Jeffrey Hancock, 'Experimental Evidence of Massive-Scale Emotional Contagion Through Social Networks' (2014) 111 Proceedings of the National Academy of Sciences of the USA 8788.
- Dietmar Schabus and others, 'One Million Posts: A Data Set of German Online Discussions' in Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (Association for Computing Machinery 2017).
- Heleen L Janssen, 'An Approach for a Fundamental Rights Impact Assessment to Automated Decision-Making' (2020) 10 International Data Privacy Law 76.
- Bertie Vidgen and others, 'Challenges and Frontiers in Abusive Content Detection' in Proceedings of the Third Workshop on Abusive Language Online (Association for Computational Linguistics 2019).
- Mike S. Schäfer, 'The Digital Public Sphere', in Gianpietro Mazzoleni et al.(eds.), *The International Encyclopedia of Political Communication* (Wiley Blackwell 2015) 322.

Policy documents, studies and contributions

- The Technology Policy Institute, "'TikTok Public Policy's Michael Beckerman" (Two Think Minimum)' (2020) <<https://techpolicyinstitute.org/2020/08/05/tiktok-public-policys-michael-beckerman-two-think-minimum/>> accessed 6 August 2020.
- Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 in the Internal Market [2000] OJ L 178/1 ("EU e-Commerce Directive").
- Proposal for a regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online (COM(2018) 640, 12 September 2018).

Legislation

- The White House, 'Executive Order on Addressing the Threat Posed by TikTok' (The White House, 2020) <<https://www.whitehouse.gov/presidential-actions/executive-order-addressing-threat-posed-tiktok/>> accessed 2 October 2020.
- Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, OJ 2016 L 119/1(hereinafter EU GDPR)
- Digital Millennium Copyright Act (USA) 17 U.S. Code § 512 (hereinafter "DMCA").
- Communications Act of 1934 at 47 U.S.C. § 230 (promulgated under the Telecommunications Act of 1996).

Arts 14 and 15 of the Internet User Public Account Information Services Management Provisions issued by Cyberspace Administration of China (Promulgated in 2017). See the English translation of the document at <https://www.chinalawtranslate.com/en/public-accounts/>.

Otto Preminger-Institut v Austria [1994] App. No. 13470/87, para 49

Marina v Romania [2020] App. No. 50469/14, para 74

Gündüz v Turkey [2003] App. No. 25071/97, para 37

Directive 2019/790/EU of the European Parliament and of the Council of 17 April 2019 and amending Directives 96/9/EC and 2001/29/EC [2019] OJ L 130/92 (“EU C-DSM Directive”).

Civil Code of the People’s Republic of China (“the Chinese Civil Code”, CCC).

Chinese Cybersecurity Law at

<https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-cybersecurity-law-peoples-republic-china/>.

German Network Enforcement Act [Netzwerkdurchsetzungsgesetz] BGBl. I S. 3352 (2017) updated as BGBl. I S. 1328, 1360 (2020).

Other

Chen Du, ‘ByteDance Cuts Domestic Engineers’ Data Access to TikTok, Other Overseas Products’ (PingWest, 2020) <<https://en.pingwest.com/a/6875>> accessed 14 June 2020.

Vanessa Pappas, ‘Introducing the TikTok Content Advisory Council’ (Newsroom | TikTok, 16 August 2019) <<https://newsroom.tiktok.com/en-us/introducing-the-tiktok-content-advisory-council>> accessed 2 October 2020

Ben Thompson, ‘The TikTok War’ (Stratechery by Ben Thompson, 14 July 2020) <<https://stratechery.com/2020/the-tiktok-war/>> accessed 15 July 2020.

Facebook Community Standards at <https://www.facebook.com/communitystandards/>.

Svea Windwehr and Jillian C. York, ‘Thank You For Your Transparency Report, Here’s Everything That’s Missing’ (Electronic Frontier Foundation, 13 October 2020) <<https://www.eff.org/deeplinks/2020/10/thank-you-your-transparency-report-heres-everything-thats-missing>> accessed 30 October 2020.

Olivia Solon, ‘Facebook’s Failure in Myanmar is the Work of a Blundering Toddler’ (the Guardian, 16 August 2018) <<http://www.theguardian.com/technology/2018/aug/16/facebook-myanmar-failure-blundering-toddler>> accessed 25 October 2020.

Aleksandra Kuczerawy, ‘The EU Commission on Voluntary Monitoring: Good Samaritan 2.0 or Good Samaritan 0.5?’ (CITIP blog, 24 April 2018) <<https://www.law.kuleuven.be/citip/blog/the-eu-commission-on-voluntary-monitoring-good-samaritan-2-0-or-good-samaritan-0-5/>> accessed 2 October 2020.

‘Number of Monthly Active Facebook Users Worldwide as of 2nd Quarter 2020’ (Statista, July 2020) <<https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/>> accessed 26 October 2020.