Disclosure Rules for Algorithmic Content Moderation

A call for a multi-level transparency regime for social media platforms

AUTHORS

Aline Iramina, University of Brasilia, Brazil, & University of Glasgow, United Kingdom
Charlotte Spencer-Smith, University of Salzburg, Austria
Wai Yan, Koe Koe Tech, Myanmar

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

Dominant social media platforms are increasingly using automation and AI to find and remove problematic content. While this helps stop some of the worst content from spreading on the Internet, algorithmic content moderation can delete content that should not be deleted ("overblocking") or discriminate against minorities. Crucially, there is very little transparency from platforms about how algorithmic content moderation works, how accurate its technologies are believed to be, and how much content they remove, especially without human review. As the use of technologies is likely to only increase, regulators should take the initiative on transparency by requiring platforms to make disclosures.

Increasing Transparency

In order to increase transparency from social media platforms, this policy brief recommends:

Adopting binding and specific disclosure rules on multiple levels for social media platforms in the use of algorithmic content moderation systems, allowing external oversight and multi-level accountability;

Requiring the implementation of robust and accessible content moderation appeal systems by social media platforms that allow users to appeal against any platforms' decision on their content, to demand human review if they were subject to an automated individual decision-making, and to quickly reinstate any legitimate content that was wrongly removed in content moderation; and

Establishing a regulatory regime that involves a multi-stakeholder approach in the rulemaking process, ensuring a more effective and efficient implementation of disclosure rules.

What is the problem?

Internet platforms, such as Facebook, Twitter, YouTube, WeChat, and TikTok, enable exciting opportunities for expression, but can also be sites of online harms, such as the posting of child abuse imagery and terrorist propaganda, the spread of hate speech and disinformation, and the facilitation of bullying and abusive activity. To counter such online harms, platforms identify problematic content or behaviour and respond by deleting or restricting it, in a process known as content moderation. Alongside human reviewers called "content moderators", platforms use automation and AI to identify and respond to problematic content and behaviour. The benefit of algorithmic content moderation (ACM) is that it is a fast and globally scalable way to prevent offensive content being uploaded and travelling across the globe within seconds. It can also spare human content moderators some of the tedium of a very repetitive job, as well as the trauma of viewing the most distressing content, such as child abuse imagery.

What is ACM?

Platforms use human reviewers, known as content moderators, to screen posts and accounts for abuse. Because of the large amount of activity that happens on platforms everyday, content moderation is too large a task for human content moderators alone. Platforms therefore use technical automation to identify and sanction violating posts and accounts. ACM refers to "systems that classify user-generated content based on either matching or prediction, leading to a decision and governance outcome (e.g. removal, geoblocking, account takedown)"[1]. This could be as simple as a bot that deletes posts with a certain keyword in them. However, large platforms routinely and increasingly use complex, advanced technologies, such as machine learning (ML), to undertake tasks in content moderation.

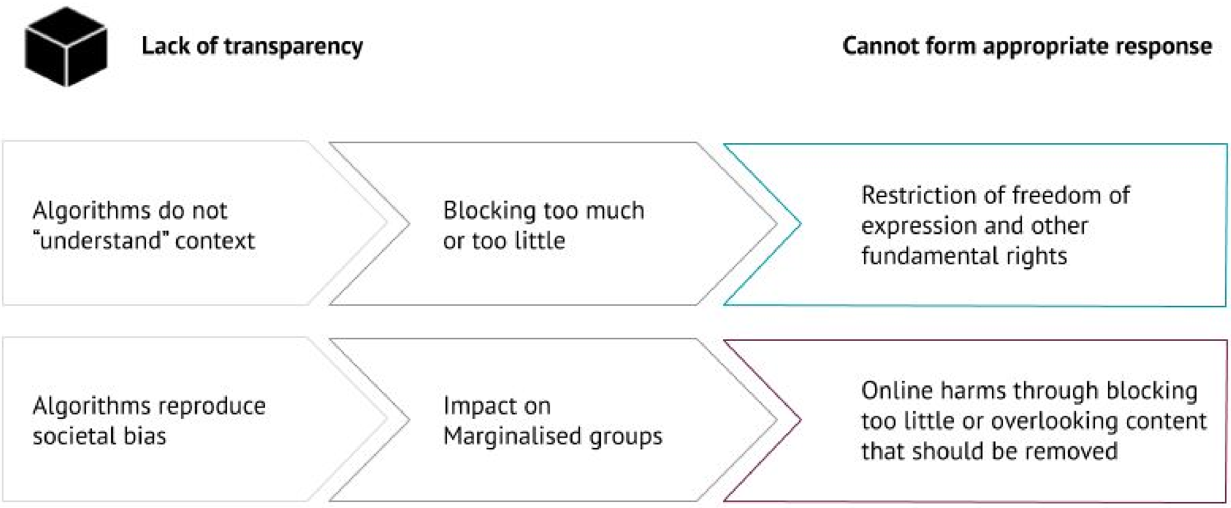
ACM, however, comes with its own risks. It does not understand context and can either block too much content or too little. "Overblocking" restricts free expression and creativity on the Internet, and can unfairly punish marginalised groups (for example, by misinterpreting certain dialects or vernacular as hate speech)[2]. Alternatively, ACM can miss abuses that would otherwise be picked up by human reviewers. Although content moderation algorithms play a central role in shaping public discourse and have a concrete impact in the "real" world, platforms release very limited information about them to the public and governments.

A lack of transparency prevents public oversight and obstructs regulators from formulating appropriate responses that would protect individual rights and public interests. Furthermore, a lack of transparency produces elevated expectations about the capabilities of algorithms as a "magic wand" for content moderation, while misleading regulators about what kinds of measures can be technically implemented, and which not. We should not expect more transparency to be a "magic wand" in itself, but it is an important first step towards appropriate approaches for future regulation and accountability mechanisms. Major platforms are planning to use more automation and AI in the future, so this issue will become ever more important.

[1] Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1). Retrieved from https://doi.org/10.1177/2053951719897945

[2] Ibid.

WHAT'S THE PROBLEM?



Lack of transparency

Cannot form appropriate response

Algorithms do not "understand" context

Blocking too much or too little

Restriction of freedom of expression and other fundamental rights

Algorithms reproduce societal bias

Impact on Marginalised groups

Online harms through blocking too little or overlooking content that should be removed

What do platforms disclose about their algorithmic content moderation practices?

Platforms use a range of different technologies in content moderation, from simple bots that delete content with certain phrases to complex ML algorithms that teach themselves which characteristics to look for when scanning content. Platforms are particularly open about the fact that they use matching technology to block content, such as child abuse imagery and terrorist propaganda. However, they are less open about how they automatically identify and remove other kinds of content which are potentially harmful, but are not illegal.

What are the main technologies employed in ACM?

Matching technologies aim to detect files that are the same as a file that has already been uploaded. Hashing is a technology used to produce a fingerprint (known as a hash) of a multimedia file, which is then matched against a collection of hashes in a database. For example, there are databases of hashed images that have already been identified as child sexual abuse imagery. Images uploaded on platforms can be compared to hashes in this database. If an abusive image is re-uploaded on social media, it can be automatically detected, without a human having to review it.

Classification technologies aim to find new violations in uploaded content by looking for patterns, e.g. in text or images. Based on pattern identification, the algorithms classify the content into predefined categories, such as nudity/not nudity. This can involve a range of different technologies, such as rules-based AI or ML. For example, platforms can look for hate speech by identifying certain keywords. In a further example, algorithms look for patterns in shapes and colours in an image that indicate that it could contain nudity[3].

[3] Ibid.

Platforms publish limited information about the use of ACM on an ad hoc basis, primarily through quarterly or biannual transparency reports about content moderation as a whole. The quality of information contained in these reports varies greatly, especially when it comes to the use of automation and AI. For example, the YouTube transparency report publishes the number of videos that are automatically taken down without human review, while the Twitter transparency report only contains overall numbers for deleted posts without differentiating between automation and human reviews (see Appendix 2 for more about the information contained in transparency reports). Platforms also publish ad hoc information on their corporate blogs and press releases, but this information is often superficial and difficult to find. Another means is through audit reports by external experts that have been commissioned by the platform, such as the Facebook Data Transparency Advisory Group. However, these reports do not give much detailed information and are intended to give recommendations to the platforms, and not to regulators. This way of sharing information does not fit the needs of the general public or regulators.

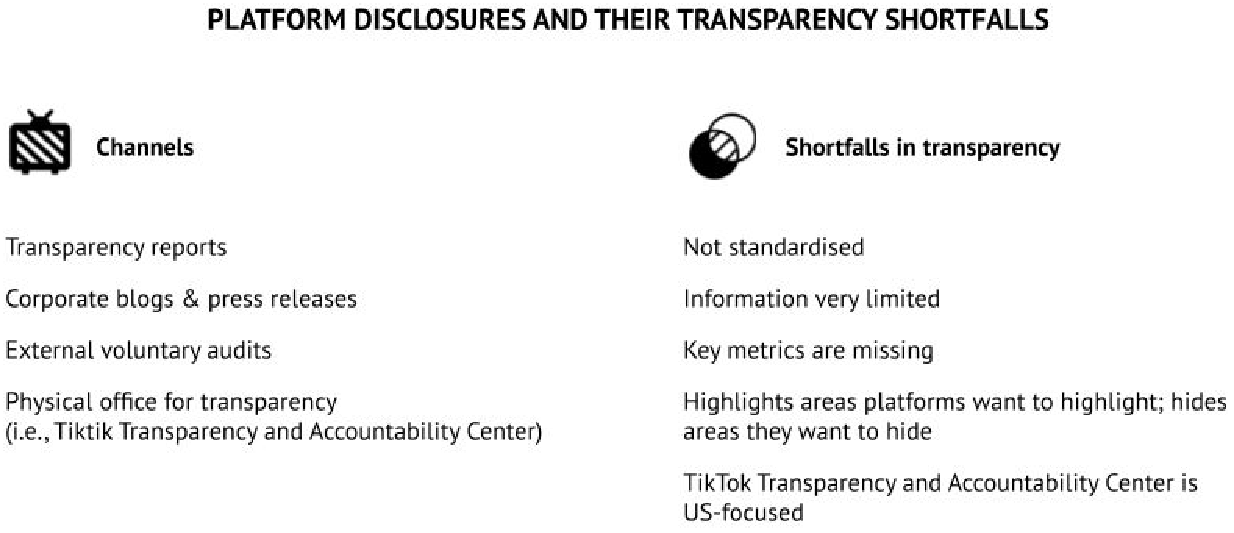
In 2020, TikTok opened a Transparency Center at their headquarters in Los Angeles, USA. The opening has been delayed because of the COVID-19 crisis, so they are offering virtual tours to journalists and experts, including a look at "[their] safety classifiers and deep learning models that work to proactively identify harmful content and [their] decision engine that ranks potentially violating content to help moderation teams review the most urgent things first"[4]. This is a step in the right direction. It seems, however, that it will not be open to the public, and TikTok will decide which information it will share, not regulators. It also seems to be focused on the US market, so it might not give insight into content moderation in other parts of the world.

A significant barrier to transparency is that the information contained in transparency reports (or indeed other information channels) is not standardised across platforms. Platforms have not yet provided direct information about how accurate algorithms are, and very seldom publish how much content is deleted automatically without human review. It is therefore difficult for regulators and the public to understand how much content is being erroneously deleted or being deleted without human intervention. Often, facts are hidden behind obscure metrics, such as Facebook's "proactive rate", which describes the percentage of content actioned that was discovered before it was reported by a user. This could mean the percentage of content actioned that was identified by algorithms, but it could also include other platform initiatives to find content, so it is not clear.

Regulators should be aware that platforms use transparency reports to draw attention towards certain topics and away from others. For example, by disclosing numbers of government information and takedown requests, platforms draw attention to the problem of government surveillance and censorship. This issue is deserving of attention, but at the same time, it also draws attention away from the role of platform companies in online speech, including the influence of proprietary content moderation technologies[5].

[4] Beckman, M. (2020). An update on our virtual Transparency & Accountability Center experience. Retrieved from https://newsroom.tiktok.com/en-us/an-update-on-our-virtual-transparency-and-accountability-center-experience

[5] Flyverbom, M. (2016). Transparency: Mediation and the Management of Visibilities. International Journal Of Communication, 10(13). Retrieved from https://ijoc.org/index.php/ijoc/article/view/4490/1531



PLATFORM DISCLOSURES AND THEIR TRANSPARENCY SHORTFALLS

Channels	Shortfalls in transparency
Transparency reports	Not standardised
Corporate blogs & press releases	Information very limited
External voluntary audits	Key metrics are missing
Physical office for transparency (i.e., Tiktik Transparency and Accountability Center)	Highlights areas platforms want to highlight; hides areas they want to hide
	TikTok Transparency and Accountability Center is US-focused

How do current legislative initiatives fall short of ensuring meaningful transparency from social media platforms about the use of algorithmic content moderation systems?

Most Internet and platform regulations currently in force do not include provisions that guarantee effective transparency and accountability from digital platforms in the use of their content moderation systems (human or automated). This is particularly true of laws that provide for limitations on the liability of digital platforms for third-party content. Much of this legislation was introduced around 20 years ago (e.g. US CDA 230, and EU E-Commerce Directive), when the use of automated content moderation was still in its infancy. Nonetheless, even more recent laws (e.g. the Brazilian Civil Rights Framework for the Internet) do not directly address the growing importance of new technologies employed by platforms for identifying and filtering illegal content, such as AI.

A great part of the national AI strategies that have been recently adopted or are currently under discussion in different countries highlights the importance of ethics of AI[6], including the importance of adopting legal frameworks that ensure that AI applications can be transparent, predictable, and verifiable. However, the existing legislation on platform regulation, as a rule, still does not take into account the risks involved in the use of AI systems for public discourse, freedom of speech, and other fundamental rights. It also treats content moderation by human and automated systems as a single entity, without differentiating them. As a consequence, although some of them provide for general transparency requirements, such as the obligation to publish transparency reports, they fall short in not providing for explicitly mandatory transparency mechanisms in the use of ACM systems, as we can see from some recent national laws and bills that introduce regulations on the topic (see Appendix 1).

In this context, the EU General Data Protection Regulation (GDPR) has already signaled an important change in the thinking around the regulation of automated systems. The GDPR requires the "data controller" to inform "data subjects" when they have been subject to fully automated individual decision-making, which is possible only in the specific cases provided for in article 22. It also introduces mechanisms that allow them to request human review or challenge the decision. Alongside the right to

[6] Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. and Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. Retrieved from http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420

know how platforms handle their personal data, users should also have the right to know how platforms handle their content, especially if they remove it. Now, with the Digital Services Act (DSA), currently under discussion in the European Commission, European regulators seem to be looking for solutions to ensure more transparency and accountability in the use of ACM systems by platforms and set the standards for other countries, as they did with the GDPR.

How should governments ensure more transparency from platforms about the use of algorithmic content moderations systems?

It is important to make clear that regulating ACM does not mean legally requiring the use of automated systems. The idea here is that regulators should be aware that algorithmic tools, although necessary, have substantive flaws and often make mistakes, and that platforms, despite their transparency efforts, are not clear about that. Therefore, it is necessary to implement rules that guarantee more transparency in platforms' decision-making process in content moderation, so it can be subjected to public argumentation and contestation[7]. Moreover, governments should have access to this information, so they can make informed decisions in their regulatory responses to the problems that might emerge in the process. For that, specific and binding disclosure rules should be the first step for more transparency.

This does not mean, however, that all data to be disclosed by platforms should necessarily be provided for in law, especially in light of the technical aspects involved in the use of algorithmic systems and their rapid technological development. While lawmakers should work on provisions that clearly define platforms' transparency obligations and explicitly enforce data disclosure on the use of these systems, with legal safeguards against the undermining of privacy, freedom of speech, and other fundamental rights, they should also leave room for further regulation by the relevant public authority with appropriate technical and policy expertise.

This public authority could be an existing authority, preferably an independent public authority with key transparency and due processes obligations[8], that would absorb these new competences related to the implementation of binding disclosure rules. However, it should work in a multi-stakeholder arrangement, with a committee, panel, or group composed of interested parties, which would be responsible for providing assistance and advice in the formulation of common technical standards (e.g. metrics) and statutory instruments that clearly establish how and what kind of data should be disclosed by platforms. This would guarantee more flexibility and technical expertise in the implementation of disclosure rules. That is why a legal requirement for the implementation of content moderation appeal systems by platforms and the establishment of a regulatory regime that involves a multi-stakeholder approach in the rulemaking process are also recommended in this policy brief. The increasing deployment of ACM should be accompanied by governmental oversight, accountability, and appeal mechanisms[9].

Adoption of multi-level disclosure rules

The adoption of multi-level specific and binding disclosure rules is an important mechanism for increasing transparency, without excessively intervening in platforms' business models. Mandatory

[7] Douek, E. (2020). Governing Online Speech: From 'Posts-As-Trumps' to Proportionality and Probability. Columbia Law Review, 121(1). Retrieved from http://dx.doi.org/10.2139/ssrn.3679607

[8] ARTICLE 19. Article 19's recommendation for the EU Digital Services Act. Retrieved from https://www.article19.org/wp-content/uploads/2020/04/ARTICLE-19s-Recommendations-for-the-EU-Digital-Services-Act-FINAL.pdf

[9] Mozilla Foundation (2020). Digital Services Act package: open public consultation. Retrieved from https://ffp4g1ylyit3jdyti1hqcvtb-wpengine.netdna-ssl.com/netpolicy/files/2020/09/Contribution1ad6cb23-a986-4b8c-9a08-80d11e5b0d47.pdf

transparency reports with only general provisions on what should be disclosed have proven insufficient. Without specific and binding disclosure obligations, platforms decide what data and information to disclose and how they will disclose them, which, in many ways, obstruct independent studies and external oversight. The option to provide for disclosure obligations on multiple levels allows governments to share oversight responsibilities with different actors (e.g. civil society, academia, and users), opting for a multi-level accountability regime[10]. Independent researchers should be able to have access to data that allow them to audit algorithms and undertake impact assessments of these algorithms, e.g. for public discourse and freedom of speech. Suggestions for multi-level disclosure include[11]:

User-facing disclosure: In addition to notifying the user of a takedown and offering the option to appeal, platforms should be required to notify users if the takedown is a result of an automated decision without human review.

Civil society and research disclosure: Platforms should be required to allow researchers access to algorithmic tools for the purposes of algorithmic auditing[12] and archived databases of deleted content, along with records of their efforts to tackle that content[13]. This should be available for most categories of account, such as copyright infringement, hate speech, and disinformation. For certain categories, such as child sexual abuse imagery and image-based abuse ("revenge porn"), there are ethical arguments against providing access. Such considerations could be negotiated in a co-regulatory structure.

General disclosure: Transparency standards should be imposed on large commercial platforms requiring information to be disclosed in a standardised form, while also allowing for differences between platforms (e.g. differences in community guidelines, or what counts as a violation). This includes provisions that oblige platforms to incorporate in their transparency reports particular information, such as:

- Explanation of how automated detection is used for each category of rule violated[14], what types of tools they are using and for what purposes, and what automated decision making approach each tool uses (e.g. natural language processing, hashing, etc.).
- Numbers of posts and accounts flagged by algorithms for each category, broken down by country[15].

[10] Leerssen, P. (2020). The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems. Retrieved from http://dx.doi.org/10.2139/ssrn.3544009

[11] In its suggestion for tiered disclosure, this paper draws significantly upon Paddy Leerssen's (2020) suggestion for tiered disclosures in social media recommender systems. Although closely interconnected, regulators should take care to differentiate between recommender systems and content moderation systems. For example, considerations of media diversity are relevant to recommender systems (e.g. are users being exposed to a sufficiently diverse array of news sources?), whereas this applies less to content moderation. Furthermore, publicly available tools providing insight into which content is recommended are already in existence (see algotransparency.org), whereas databases of certain categories of removed content should not be made publicly available for ethical reasons (such as child abuse imagery and image based abuse, also known as "revenge porn").

[12] Schmon, C. (2020). EFF Responds to EU Commission on the Digital Services Act: Put Users Back in Control. Retrieved from

[13] Bowers, J. & Zittrain, J. (2020). Answering impossible questions: content governance in an age of disinformation. The Harvard Kennedy School (HKS) Misinformation Review. Retrieved from https://misinforeview.hks.harvard.edu/article/content-governance-in-an-age-of-disinformation/

[14] Santa Clara Principles (2018). Santa Clara Principles. Retrieved from https://santaclaraprinciples.org; European Regulators Group for Audiovisual Media Services (2020). ERGA Position Paper on the Digital Services Act. ERGA 2020 Subgroup 1 - Enforcement. Retrieved from

[15] Ibid.

- Numbers of posts and accounts deleted automatically for each category[16], broken down by country.
- Estimated accuracy rates for flagging and how accuracy is defined.
- As content moderation is geographically patchy across the globe, platforms should also be required to disclose roll-outs of new AI systems (e.g. the AI system being rolled out in the jurisdiction and for which kinds of violations, e.g. new roll-out of AI against hate speech in a particular country).

Implementing content moderation appeal systems

The implementation of robust and accessible content moderation appeal systems is considered an essential mechanism to guarantee more accountability from platforms. It is in line with users' right to information, particularly to be notified when they are subject to automated decisions without human review, since they should also have the right, for example, to appeal against any decision they consider wrong[17]. ACM without human review entails the risk of content being wrongly removed, and as we have seen during COVID-19 crisis, its use is on the increase. Appeal mechanisms have become even more necessary to guarantee the rapid reinstatement of any legitimate content or account that was wrongly removed, mitigating potential harmful effects for public discourse, freedom of speech, and other users rights.

Establishing a multi-stakeholder-based regulatory regime

The establishment of a regulatory regime that involves a multi-stakeholder approach in the rulemaking process is also a key mechanism to guarantee more effectiveness and efficiency in the implementation, regulation, monitoring, and enforcement of disclosure rules. The relevant public authority should include in its decision-making process a multi-stakeholder group, with representatives of interested parties, such as platforms, independent researchers, users, and civil society organizations. This multi-stakeholder group should provide assistance and advice to the public authority in drafting common disclosure standards and rules. It could also assist the public authority in monitoring whether platforms are complying with their disclosure obligations in a timely, accurate, and complete manner. It could, for example, advise on cases involving non-compliance complaints against platforms and the application of sanctions in case of non-compliance. There is not one single model for the adoption of a multi-stakeholder approach[18]. This multi-stakeholder group could be part of the public authority structure or an external advisory group, depending on the existing infrastructure or the preferences of the country or region in question. Governments have already adopted this kind of structure for other regulated sectors, so they can look at them as a reference.

There are some other suggestions that also propose a multi-stakeholder approach, particularly from civil society groups. One example is the creation of an independent, accountable, and transparent multi-stakeholder body, which would be responsible for elaborating and implementing the technical and practical remedies necessary to guarantee more transparency and accountability from platforms, as set in

[16] Ibid.

[17] Santa Clara Principles (2018). Santa Clara Principles. Retrieved from https://santaclaraprinciples.org

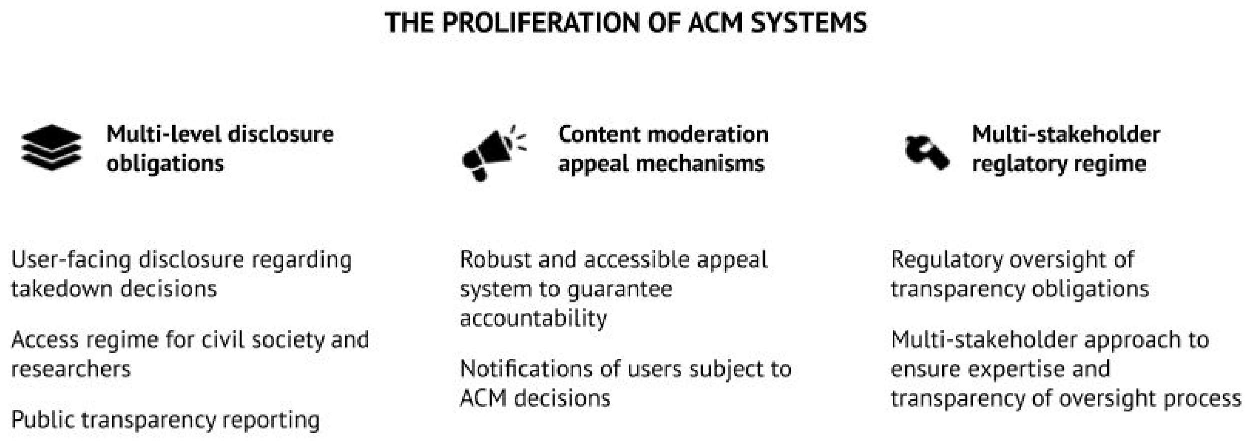
[18] "Multistakeholder approach is a set of tools or practices that all share one basis: Individuals

and organizations from different realms participating alongside each other to

share ideas or develop consensus policy". Retrieved from https://www.internetsociety.org/resources/doc/2016/internet-governance-why-the-multistakeholder-approach-works/.

the law[19]. In this case, some of the competences of the relevant public authority would be transferred to the independent multi-stakeholder body, which would be monitored by this public authority. This would also be a viable solution and, depending on the public structure and resources available in each country, using existing infrastructure and involving directly existing stakeholders could be an option to reduce some of the costs involved in the implementation of these rules by states.

In order to achieve the expected results with the adoption of these transparency and accountability mechanisms, one of the biggest challenges for governments will be to gather the knowledge and technical expertise necessary to implement them. That is why it is so important to have a multi-stakeholder approach already in the regulatory phase. The option of sharing responsibilities with other stakeholders through a multi-level accountability regime, in our view, is one of the best solutions to tackle platforms' lack of transparency in the use of ACM systems.



THE PROLIFERATION OF ACM SYSTEMS

Multi-level disclosure obligations

User-facing disclosure regarding takedown decisions

Access regime for civil society and researchers

Public transparency reporting

Content moderation appeal mechanisms

Robust and accessible appeal system to guarantee accountability

Notifications of users subject to ACM decisions

Multi-stakeholder reglatory regime

Regulatory oversight of transparency obligations

Multi-stakeholder approach to ensure expertise and transparency of oversight process

[19] ARTICLE 19. Article 19's recommendation for the EU Digital Services Act. Retrieved from https://www.article19.org/wp-content/uploads/2020/04/ARTICLE-19s-Recommendations-for-the-EU-Digital-Services-Act-FINAL.pdf

References

Articles and Research Papers

Beckman, M. (2020). *An update on our virtual Transparency & Accountability Center experience*. Retrieved from https://newsroom.tiktok.com/en-us/an-update-on-our-virtual-transparency-and-accountability-center-experience

Bowers, J. & Zittrain, J. (2020). *Answering impossible questions: content governance in an age of disinformation*. The Harvard Kennedy School (HKS) Misinformation Review. Retrieved from https://misinforeview.hks.harvard.edu/article/content-governance-in-an-age-of-disinformation/

Douek, E. (2020). Governing Online Speech: From 'Posts-As-Trumps' to Proportionality and Probability. *Columbia Law Review*, 121(1). Retrieved from http://dx.doi.org/10.2139/ssrn.3679607

Flyverbom, M. (2016). Transparency: Mediation and the Management of Visibilities. *International Journal Of Communication, 10*(13). Retrieved from https://ijoc.org/index.php/ijoc/article/view/4490/1531

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1). Retrieved from https://doi.org/10.1177/2053951719897945

Leerssen, P. (2020). *The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems*. Retrieved from http://dx.doi.org/10.2139/ssrn.3544009

Marsden, C. (2011). Internet co-regulation and constitutionalism. In *Internet Co-Regulation: European Law, Regulatory Governance and Legitimacy in Cyberspace* (pp. 46-70). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511763410.003

Policy documents, studies and contributions

AlgorithmWatch (2020). *Our response to the European Commission's planned Digital Services Act*. Retrieved from https://algorithmwatch.org/en/submission-digital-services-act-dsa/#audit

ARTICLE 19. Article 19's recommendation for the EU Digital Services Act. Retrieved from https://www.article19.org/wp-content/uploads/2020/04/ARTICLE-19s-Recommendations-for-the-EU-Digital-Services-Act-FINAL.pdf

Ausloos, J., Leerssen, P., Thije, P. (2020). *Operationalizing Research Access in Platform Governance What to learn from other industries?*. Retrieved from https://www.ivir.nl/publicaties/download/GoverningPlatforms_IViR_study_June2020-AlgorithmWatch-2020-06-24.pdf

European Regulators Group for Audiovisual Media Services (2020). *ERGA Position Paper on the Digital Services Act. ERGA 2020 Subgroup 1 - Enforcement*. Retrieved from https://erga-online.eu/wp-content/uploads/2020/06/ERGA_SG1_DSA_Position-Paper_adopted.pdf

European Parliament (2020). *Report with recommendations to the Commission on a Digital Services Act: adapting commercial and civil law rules for commercial entities operating online*. Retrieved from https://www.europarl.europa.eu/doceo/document/A-9-2020-0177_EN.html#title1

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. and Srikumar, M. (2020). *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*. Retrieved from http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420

Global Network Initiative (2020). *Digital Services Act package: open public consultation*. Retrieved from https://globalnetworkinitiative.org/wp-content/uploads/2020/09/DSA-Survey-GNI-Submission-1.pdf

Mozilla Foundation (2020). *Digital Services Act package: open public consultation*. Retrieved from https://ffp4g1ylyit3jdyti1hqcvtb-wpengine.netdna-ssl.com/netpolicy/files/2020/09/Contribution1ad6cb23-a986-4b8c-9a08-80d11e5b0d47.pdf

Intervozes, Observacom, Idec and Desarrollo Digital (2019). *Contribuições para uma Regulação Democrática das Grandes Plataformas que garanta a Liberdade de Expressão na Internet*. Retrieved from https://www.observacom.org/wp-content/uploads/2019/08/Contribuic%cc%a7o%cc%83es-para-uma-regulac%cc%a7a%cc%83o-democra%cc%81tica-das-grandes-plataformas-que-garanta-a-liberdade-de-expressa%cc%83o-na-internet.pdf

UK Home Office and UK Department for Digital, Culture, Media & Sport (2020). *Online Harms White Paper - Initial Consultation response*. Retrieved from https://www.gov.uk/government/consultations/online-harms-white-paper/public-feedback/online-harms-white-paper-initial-consultation-response#contents

Santa Clara Principles (2018). *Santa Clara Principles*. Retrieved from https://santaclaraprinciples.org

Schmon, C. (2020). *EFF Responds to EU Commission on the Digital Services Act: Put Users Back in Control*. Retrieved from https://www.eff.org/deeplinks/2020/09/eff-responds-eu-commission-digital-services-act-put-users-back-control

Legislation

Communications Decency Act of 1996 47 USC § 230

Criminal Code Amendment (sharing of abhorrent violent material) Act 2019 (AUS)

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

Federal Law on Amendments to Article 10 of Federal Law n. 149-FZ of July 27, 2006 on Information, Informational Technologies and the Protection of Information (RUS)

Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken [Netzwerksdurchsetzungsgesetz] [NetzDG]) (DEU)

Hate Speech and Disinformation Prevention and Suppression Proclamation No.1185 /2020 (ETH)

İnternet Ortamında Yapılan Yayınların Düzenlenmesi ve Bu Yayınlar Yoluyla İşlenen Suçlarla Mücadele Edilmesi Hakkında Kanun. T.C. Resmî Gazete. 23 Mayıs 2007. 3 Ağustos 2011 tarihinde kaynağından arşivlendi. Erişim tarihi: 29 Temmuz 2020 (TUR)

Lei n° 12965 de 23 de abril de 2014 estabelece princípios garantias direitos e deveres para o uso da Internet no Brasil (BRA)

Protection from Online Falsehoods and Manipulation Act (No. 18 of 2019) (SGP)

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 [2016] OJ L 119/1

Appendices

Appendix 1

Table 1. (Laws and Bills with legal provisions on content moderation by social media platforms)

		Moderated Content**			Mandatory transparency measures (e.g. transparency reports)	AI and algorithmi transparency measures***
	Law/Bill*	Hate Speech	Fake News	Other harmful or illegal content		
Australia	Criminal Code Amendment (sharing of abhorrent violent material)	x		x		
Austria	Draft on measures to protect users on communication platforms (Communications Platforms Act)	x		x	x	
Brazil	Brazilian Law on Freedom, Responsibility and Transparency on the Internet (Bill No. 2630/2020)	x	x	x	x	
Ethiopia	Hate Speech and Disinformation Prevention and Suppression Proclamation No.1185 /2020	x	x			
Germany	German Social Networks Enforcement Act (NetzDG)	x	x	x	x	
India	The Information Technology Intermediaries Guidelines (amendment) rules, 2018	x	x	x		
Russia	Federal Law on Amendments to Article 10 of Federal Law on Information, Information Technologies and the Protection of Information	x	x	x	x	
Singapore	Protection from Online Falsehoods and Manipulation Act (No. 18 of 2019)		x			
Turkey	Law No. 7253 amending "Law No. 5651 on The Regulation of Publications made in the Internet Environment and Combatting Crimes Committed through these Publications."	x	x	x	x	

*National laws introduced in the last 4 years and bills currently under discussion that introduce regulation on content moderation by social media platforms.

** There are some cases where the content addressed by the law is not clear, particularly in relation to fake news.

***Transparency measures in the use of AI and algorithms for content moderation, not for other purposes. Some laws or bills include among the information to be provided by platforms, for example, methods or methodology employed in the detection of irregularity, which could include information on algorithms and AI, but it is not clear about that.

Appendix 2

Table 2. Information disclosed about ACM in transparency reports (based on most recent reports as of October 2020)

Platform	Information disclosed
Facebook & Instagram	● "Proactive rate": percentage of content actioned that was found before users report it. This includes "detection technology", but it is vague as to whether the "proactive rate" means content removed automatically, or if content removed by other means is also included. ● Appeal rates: content deletions appealed by the user. This cannot be taken as a proxy for accuracy, because users do not necessarily appeal when content is wrongfully removed. ● Reinstatement rates: removed content that is later restored, either as a result of appeal or because Facebook has independently decided to restore it. This also cannot be taken as a proxy for accuracy, because the number depends, in part, on the appeal rate. ● Rates are given by content category, but not by country. ● Occasionally, blog posts that accompany the release of a new transparency report will make mention of new developments in technology, such as the roll-out of a new AI in a geographic region.
Reddit	● Content moderation relies mostly on Reddit admins, who set the Content Policy, and community moderators, who set Community Rules specific to a community called a subreddit. Reddit also has a tool called AutoModerator, which can be customized by the moderator to set rules and help them in moderation, e.g. reporting or removing comments containing certain phrases or links. ● The Reddit transparency report 2019 publishes overall content removals and appeals as well as government requests for content removal and user information. Regarding ACM, it only mentions that they use PhotoDNA and YouTube CSAI Match to detect child sexual abuse imagery.
YouTube (Google)	● YouTube reports the number of videos removed by automated flagging. However, this only includes videos removed because they violate YouTube's own rules, and does not include videos that are deemed to be illegal in various countries. Numbers for copyright infringement are not reported. ● Appeal and reinstatement rates are reported, but these cannot be seen as a proxy for accuracy rates. ● Rates are given by content category and by country, but only for overall takedowns.
Twitter	● Twitter does not publish information about automation, algorithms, or AI in its content moderation transparency reports. Instead, it publishes overall numbers of accounts actioned, content removed, and accounts suspended.
TikTok	● TikTok gives the total global number of videos flagged and removed automatically for violating their guidelines. This number is not broken down by country or category. ● TikTok breaks down numbers per category and country, but only for overall takedowns (including means other than ACM). It also excludes the number of automatic removals from its country breakdowns.