

Some human and technical aspects of online content regulation

Gosztonyi, Gergely

Veröffentlichungsversion / Published Version

Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:

Gosztonyi, G. (2021). Some human and technical aspects of online content regulation. *Journal of Liberty and International Affairs*, 7(Supp. 1), 149-159. <https://doi.org/10.47305/JLIA21371149g>

Nutzungsbedingungen:

Dieser Text wird unter einer CC BY Lizenz (Namensnennung) zur Verfügung gestellt. Nähere Auskünfte zu den CC-Lizenzen finden Sie hier:

<https://creativecommons.org/licenses/by/3.0/deed.de>

Terms of use:

This document is made available under a CC BY Licence (Attribution). For more information see:

<https://creativecommons.org/licenses/by/3.0>

Copyright © 2021 The Author/s
This work is licensed under a CC-BY 3.0 License
Peer review method: Double-Blind
Accepted: September 16, 2021
Published: November 23, 2021
Review article
DOI: <https://www.doi.org/10.47305/JLIA21371149g>

SOME HUMAN AND TECHNICAL ASPECTS OF ONLINE CONTENT REGULATION

Gergely Gosztonyi

Eötvös Loránd University (ELTE), Faculty of Law, Hungary
ORCID iD: <https://orcid.org/0000-0002-6551-1536>
gosztonyi@ajk.elte.hu

Abstract: The amount of newly uploaded content on the internet is growing daily: 60 seconds on the web in 2021 consist of more than 500 hours of content uploaded on YouTube, 695,000 stories shared on Instagram, and nearly 70 million messages sent via WhatsApp and Facebook Messenger. The vast majority of them is legal content, but a slice is illegal or harmful. The article analyses the situation and the problems of both human and AI moderation, then it gives an answer how to handle the content on the internet with a shared usage of human and AI moderation as they could perfectly complement each other in a long term.

Keywords: Artificial Intelligence; Moderation; Content Regulation; Facebook; Shared Usage

INTRODUCTION

Between 1612 and 1614, the world-famous Flemish painter Peter Paul Rubens painted his painting 'Deposition from the Cross' and around 1614-15 he painted 'Venus in front of the Mirror'. It was probably not only because of the obscurity of social media at the time that he never imagined that his painting – or, more precisely, the events surrounding it – would make a company an object of ridicule a few centuries later. In 2018, the state-owned Flemish Tourist Office launched an advertising campaign on Facebook using these two images, but Facebook removed both paintings because they contained gratuitous nudity. In a letter to the company, the Office wrote: "Even though we secretly have to laugh about it your cultural censorship is making life rather difficult for us" (Frieze 2018).¹ Previously Gustave Courbet's 1866 painting 'The Origin of the World' (which was the subject of an eight-year court case in France²) or the photo of

¹ It should be added, however, that the events saw a fortunate increase in traffic to Flemish museums during the period, and the office produced a video mocking Facebook's policy called 'Social media doesn't want you to see Rubens' paintings', <https://www.youtube.com/watch?v=UZq3cVgU5AI>

² For the interesting case that ended in a settlement between the parties, see: Cascone 2019.

Venus de Willendorf met the same fate (Dawson 2018), as the explicit depiction of nude body parts was deemed pornography by the company's Community Standards (CS) moderation guidelines. But it's not just paintings that have faced a similar situation: the iconic statue of the Little Mermaid in the Danish capital, Copenhagen was also removed because it contained too much bare skin or sexual undertones. A painting of Santa Claus kneeling before baby Jesus in a manger has also been removed for violent content (Stagnaro 2018). But it's also easy to imagine being banned for a photo of a cat: a user had his account blocked in 2016 for sharing a picture of a cat in a suit. The reasons were unknown to the public (Moore 2016). Also, at the request of the United States' police forces, some content in which a black woman took a video of white police officers has been taken down (Karr 2016). The list goes on and on.

For all these reasons, and as a result of the campaign by the US NGO 'National Coalition Against Censorship', Facebook agreed that CS needed to be reviewed, but the process was time-consuming.³ As Daily Mail genuinely pointed out: "after all, a butt is a butt and a nipple is a nipple. But deciding when a nipple is an art, porn or protest gets murky even when humans are doing the deciding. Teaching AI software about human sexual desire is a whole other ballgame" (Daily Mail Online 2020).

But why should be all these examples interesting to all of us?

Facebook had 2.8 billion users in 2020 and the company "generates 4 petabytes of data per day – that's a million gigabytes. All that data is stored in what is known as the Hive, which contains about 300 petabytes of data" (Roy 2020). Whilst, "60 seconds on the web in 2021 consist of more than 500 hours of content uploaded on YouTube, 695,000 stories shared on Instagram and nearly 70 million messages sent via WhatsApp and Facebook Messenger" (Jenik 2021). Much of this vast amount of data is legitimate content, but even the small amount that shouldn't be there is still huge. The inspection of this content should be handled somehow.

FACEBOOK MODERATION GUIDELINES OR CONTENT IN QUESTION

In the preamble to the moderation policy, the company states: "We take our role seriously in keeping abuse off the service. (...) The goal of our Community Standards is to create a place for expression and give people a voice".⁴ The Community Standards (CS) identifies four core values that it follows and expects all its users to adhere to authenticity, security, privacy, and dignity. The CS has five categories for the types of content it does not wish to host on its platform (Facebook 2021):

a) Violence and crime:

³ Angelo Stagnaro probably misunderstood the process as he writes: "Facebook has had a long history of censoring Christian organizations and individuals, flagging our beliefs as being 'hateful' or otherwise inappropriate, but it is their actions and inaction that is most accurately branded as hateful" (Stagnaro 2018).

⁴ Facebook 2021.

- a. Violence and incitement;
 - b. Dangerous individuals and organizations;
 - c. Coordinating harm and publicizing crime;
 - d. Regulated goods;
 - e. Fraud and deception.
- b) Security:
- a. Suicide and self-injury;
 - b. Child sexual exploitation, abuse, and nudity;
 - c. Sexual exploitation of adults;
 - d. Bullying and harassment;
 - e. Human exploitation;
 - f. Privacy violations and image privacy rights.
- c) Objectionable content:
- a. Hate speech;
 - b. Violent and graphic content;
 - c. Adult nudity and sexual activity;
 - d. Sexual solicitation.
- d) Integrity and credibility:
- a. Spam;
 - b. False news;
 - c. Manipulated media.
- e) Respect for intellectual property:
- a. Intellectual property infringement.

These are the types of content where the legislator often encounters difficulties. However, it should be pointed out that in the era of 'privatization' of content regulation (Hinzt 2015), these issues were the responsibility of the service providers only, so it is not particularly surprising that CS is constantly changing. In 2018, after a long time and many complaints, Facebook made its moderation policies public, as Mónica Pintér (2018) put it so eloquently: "if Facebook is a country, here is its new constitution".⁵ The big change was not just the publicity, but also the fact that the internal policies that govern the company's content regulation practices were made public. This means that content is classified using a combination of artificial intelligence (AI) and human intervention: together they are the first line of defense in the war against unsolicited

⁵ Cf. the term 'Facebookistan' coined by Rebecca MacKinnon (2013).

content. But why only the first line of defense, one may ask. Tarleton Gillespie (2018) in his book: 'Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media', compared the actors involved in moderation to a pyramid, each with their role to play:

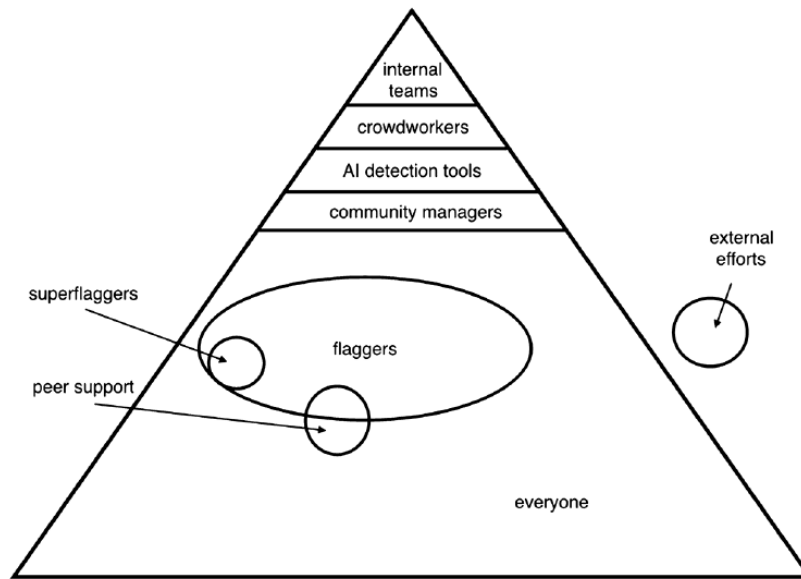


Figure 1: Pyramid of participants in content management (Source: Gillespie 2018)

MODERATION USING ARTIFICIAL INTELLIGENCE

Artificial intelligence⁶ in its current form still has many difficulties to overcome context, irony, slang, etc. It "is very good at identifying porn, spam, and fake accounts, but it's still not great at identifying hate speech" (Koebler and Cox 2018). Users create content in over a hundred languages and, in its current form, AI is not yet capable of interpreting more complex content. But the need is obvious: there is not enough human workload that could handle the amount of data in the 2020s. Facebook „says its AI tools – many of which are trained with data from its human moderation team – detect nearly 100 percent of spam, and that 99.5 percent of terrorist-related removals, 98.5 percent of fake accounts, 96 percent of adult nudity and sexual activity, and 86 percent of graphic violence-related removals are detected by AI, not users" (Koebler and Cox 2018).

In contrast to all these figures, AI was only able to correctly detect, interpret and manage 36 percent of hate speech content.⁷ Concerning the process, Thiago Dias Oliva

⁶ It should be stressed that although the term 'artificial intelligence' is used in public discourse, it "can refer to the use of a variety of automated processes at different phases of content moderation" (Llansó *et al.* 2020).

⁷ Joaquin Quiñonero Candela, a Director of AI at Facebook was interviewed about the shortcomings of Facebook and its use of AI (Hao 2021).

also points out that “whilst traditional law enforcement encompasses detection, prosecution, adjudication and punishment performed by different actors, algorithmic content policing does that all at once, focusing on early detection and prevention in a less transparent fashion” (Dias Oliva 2020). In April 2021, the United States Federal Trade Commission issued a statement warning that there are also worrying racial and gender biases associated with AI (Jillson 2021). Kinga Sorbán (2021) also points to the dangers of automated filtering systems and freedom of speech. In addition to all this, Jack M. Balkin ironically points to the economic rationality that “algorithmic employees cost even less than human employees: they do not have families, they do not take coffee breaks” (Balkin 2018).

Emma Llansó *et al.* (2020) summarised the legislative dilemmas related to the use of AI as follows:

- In the public discourse, moderation by artificial intelligence should be replaced by something else that takes into account a broader range of automated technologies and processes.
- Automation in content moderation should not be mandated in law because the state of the art is neither reliable nor effective.
- Tech companies using automatic moderation should provide more transparency on their procedures.
- Not everything can be solved by automatic moderation.
- As automatic moderation can also lead to conflicts with fundamental rights – most notably the freedom of expression – it is important to ensure that there is no over removal.
- There is no ‘neutral’ automatic moderation.
- Developing media literacy is crucial to this issue.

MODERATION BY HUMANS

As the exponentially growing amount of content and number of users cannot be handled by artificial intelligence, giant tech companies have responded by hiring more and more people. The change in ten years is almost unimaginable: “at Facebook (...), in 2009, only twelve moderators were in charge of examining the content and deciding content disputes. They tried to make fair decisions on content and conflicts of law posted by the then one hundred and twenty million users” (Huszár 2021). Compare that to 2018: “Facebook employs a total of 7,500 moderators worldwide, a position that now employs 40 percent more than this time last year” (nlc.hu 2018). These workers, called Community Operations (CO), are the ones who review reports from users twenty-four hours a day, and in the vast majority of cases, they manage to process the reports within twenty-four hours. Very little is known about their work, which – companies say – is in their defense against users. The information leaked is generally terrible (post-traumatic

reactions to the content viewed, terrible working conditions, low wages, outsourced workers, constant stress and emotional pressure, mental health problems, and unimaginable fluctuation as a result of all this) (Newton 2019a)⁸, to which Mark Zuckerberg reacted as: "some of the reports, I think, are a little overdramatic" (Newton 2019b). On the positive side, however, the company has "a team of four clinical psychologists in three regions to help moderators who regularly encounter violent, abusive content" (hvg.hu 2019). And there are times when it may be needed: based on the reports and the amount of data, it takes roughly thirty seconds to decide on an entry in around 400 cases a day – how to assess and interpret the context of the text in that shortage of time is at least questionable. According to Ruckenstein and Turunen, human moderators are increasingly being treated as machines by tech companies, which will cause problems in the long term concerning the working environment (Ruckenstein and Turunen 2020).

As moderators are not experts in a particular area, Facebook provides them with a guide, with yes/no answers to a series of questions, at the end of which they can decide whether or not to remove content. Context, languages, dialects, and user intent can further complicate the issue. In addition, legislation can vary considerably from country to country, so this 'world guide' is not always useful. This brings us to Kyle Langvardt's question: are we sure that 'privatizing' the decision to regulate content is the right way forward? (Langvardt 2017).

CONCLUSION

The United Nations (UN) Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression has issued a report calling for "radical transparency" (UNHRC 2018) from social media providers in the way they make and enforce their rules. Shared use of human moderation alongside AI can filter out one of the most dangerous problems: bias. And although one might think that this is only true for human moderators, there is also the concept of the so-called "discriminatory algorithm" (Turner Lee, Resnick and Barton 2019), which is caused by insufficient data and which, as a consequence, mainly affects minority groups underrepresented in social media. YouTube has also reported, "excessive censorship" of its own AI tools (Barker and Murphy 2020).

The possible solution to moderation is summarised in a report prepared for Ofcom, the regulatory and competition authority for the broadcasting, telecommunications, and postal industries of the United Kingdom:

⁸ cf. (Buni and Chemaly 2016).

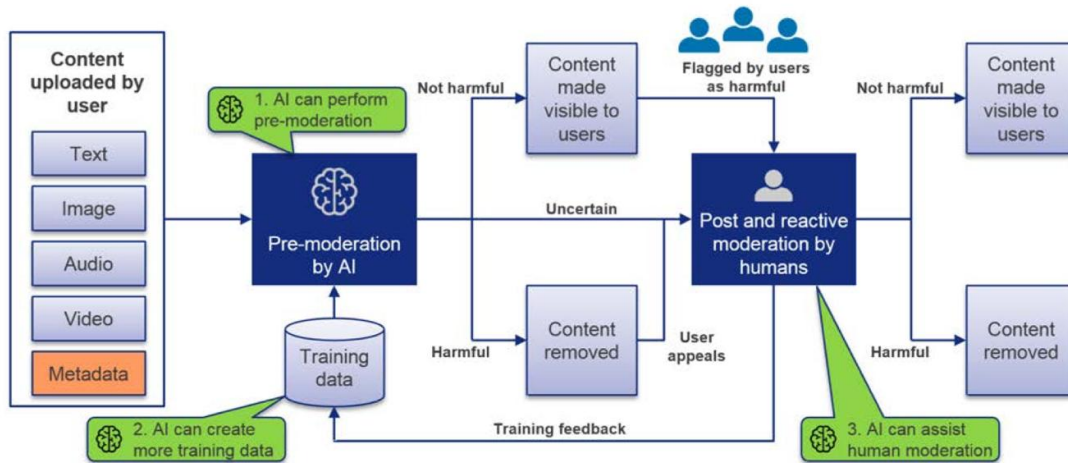



Figure 2: Ideal Flowchart for Moderation (Source: Cambridge Consultants 2019)

According to the study, the use of artificial intelligence can perfectly complement human moderation in three ways:

- AI can be used to improve the pre-moderation stage and flag content for review by humans, increasing moderation accuracy.
- AI can be implemented to synthesize training data to improve pre-moderation performance.
- AI can assist human moderators by increasing their productivity and reducing the potentially harmful effects of content moderation on individual moderators (Cambridge Consultants 2019).

The solution is therefore unlikely to be a choice between human moderation or moderation by AI, but rather a combination of the two in the future. In 2020, during the Covid-19 pandemic – while big tech companies were also requiring employees to work from home and giving artificial intelligence more tasks (Gillespie 2020) – “Facebook and Google roughly doubled the amount of potentially harmful material they removed in the second quarter of this year compared with the three months through March” (Scott and Kayali 2020), and there were many more complaints about the decisions as a result, making it clear that human content scrutiny will not be unnecessary for some time (Barker and Murphy 2020). 

COMPLIANCE WITH ETHICAL STANDARDS

Acknowledgments:

Not applicable.

Funding:

Not applicable.

Statement of human rights:

This article does not contain any studies with human participants performed by any of the authors.

Statement on the welfare of animals:

This article does not contain any studies with animals performed by any of the authors.

Informed consent:

Not applicable.

REFERENCES

1. Balkin, Jack M. 2018. „Free speech is a triangle“. *Columbia Law Review* 7: 2024.
2. Barker, Alex and Murphy, Hannah. 2020. „YouTube reverts to human moderators in the fight against misinformation“. *Financial Times*, September 20.
<https://www.ft.com/content/e54737c5-8488-4e66-b087-d1ad426ac9fa>
3. Buni, Catherine and Chemaly, Soraya. 2016. „The secret rules of the internet. The murky history of moderation, and how it's shaping the future of free speech“. *The Verge*, April 13. <https://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech>
4. Cambridge Consultants. 2019. „Use of AI in online content moderation“ *Ofcom*,
https://www.ofcom.org.uk/data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf
5. Cascone, Sarah. 2019. „After an Eight-Year Legal Battle, Facebook Ends Its Dispute With a French School Teacher Who Posted Courbet's 'Origin of the World'“. *artnet*, August 5. <https://news.artnet.com/art-world/facebook-courbet-lawsuit-ends-1616752>
6. Daily Mail Online. 2020. „Do naked bodies belong on Facebook? Tech giant struggles with changing 'vague and unevenly enforced' rules over nudity and body art without suppressing freedom of speech“ *Daily Mail Online*, January 15.
<https://www.dailymail.co.uk/sciencetech/article-7890913/Does-naked-body-belong-Facebook-It-s-complicated.html>
7. Dawson, Aimee. 2018. „Facebook censors 30,000-year-old Venus of Willendorf as 'pornographic'“. *The Art Newspaper*, February 27.
<https://www.theartnewspaper.com/news/facebook-censors-famous-30-000-year-old-nude-statue-as-pornographic>
8. Dias Oliva, Thiago. 2020. „Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression“. *Human Rights Law Review* 4: 612.
9. Facebook. 2021. „Facebook Community Standards“,
<https://transparency.fb.com/en-gb/policies/community-standards>
10. Frieze. 2018. „Facebook Censors Rubens Paintings For Nudity“ *frieze*, July 24.
<https://www.frieze.com/article/facebook-censors-rubens-paintings-nudity>
11. Gillespie, Tarleton. 2018. *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*. New Haven–London: Yale University Press, 116.
12. Gillespie, Tarleton. 2020. „Content Moderation, AI, and the Question of Scale“. *Big Data & Society* 2: 2.
13. Hao, Karen. 2021. „How Facebook got addicted to spreading misinformation“, *MIT Technology Review*, March 11,

<https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation>

14. Hintz, Arne. 2015. „Social media censorship privatized regulation, and new restrictions to protest and dissent” In *Critical Perspectives on Social Media and Protest: Between Control and Emancipation*, edited by Dencik, Lina and Leistert, Oliver. London: Rowman & Littlefield, 109–126.
15. Huszár, Daniella. 2021. „A véleménynyilvánítás és kifejezés szabadsága a közösségi média platformjain [Freedom of opinion and expression on social media platforms]” *Média Kábel Műhold* 4: 34.
16. hvg.hu. 2019. „Mégis megszólalt a Facebook: így dolgoznak a bejegyzéseinket átnéző moderátorok [Facebook has spoken: this is how moderators reviewing our posts work]” *hvg.hu*, July 29, <https://hvg.hu/tudomany/20180728facebookmoderatorbejegyzesekellenorzese>
17. Jenik, Claire. 2021. "A Minute on the Internet in 2021". *Statista.com*, July 30, <https://www.statista.com/chart/25443/estimated-amount-of-data-created-on-the-internet-in-one-minute>
18. Jillson, Elisa. 2021. „Aiming for truth, fairness, and equity in your company's use of AI” *ftc.gov*, April 19. <https://www.ftc.gov/news-events/blogs/business-blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai>
19. Karr, Timothy. 2016. „How Censoring Facebook Affects the Fight for Black Lives”. *The Root*, August 29. <https://www.theroot.com/how-censoring-facebook-affects-the-fight-for-black-live-1790856542>
20. Koebler, Jason and Cox, Joseph. 2018. „The Impossible Job: Inside Facebook's Struggle to Moderate Two Billion People”. *Vice*, August 23. <https://www.vice.com/en/article/xwk9zd/how-facebook-content-moderation-works>
21. Langvardt, Kyle. 2017. „Regulating Online Content Moderation”. *Georgetown Law Journal* 5: 1387.
22. Llansó, Emma et. al. 2020. *Artificial Intelligence, Content Moderation, and Freedom of Expression*. Transatlantic Working Group. <https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf>
23. MacKinnon, Rebecca. 2013. *Consent of the Networked. The Worldwide Struggle For Internet Freedom*. New York: Basic Books.
24. Moore, Michael. 2016. „Facebook sharing this cat photo could get you BANNED” *Daily Express*, October 5. <https://www.express.co.uk/life-style/science-technology/717978/facebook-bans-user-sharing-cat-photo>
25. Newton, Casey. 2019a. „The trauma floor. The secret lives of Facebook moderators in America”. *The Verge*, February 25. <https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona>

26. Newton, Casey. 2019b. „Read the full transcript of Mark Zuckerberg's leaked internal Facebook meetings”. *The Verge*, October 1.
<https://www.theverge.com/2019/10/1/20892354/mark-zuckerberg-full-transcript-leaked-facebook-meetings>
27. nlc.hu. 2018. „Nyilvánosak a Facebook moderálási alapelvei [Facebook's moderation principles are public]” *nlc.hu*, April 24.
<https://nlc.hu/szabadido/20180424/facebook-moderalasi-alapelvek-nyilvanos-fellebbezes>
28. Pintér, Mónika. 2018. „Ha a Facebook egy ország, itt az új alkotmánya [If Facebook is a country, here is its new constitution]” *24.hu*, April 24.
<https://24.hu/tech/2018/04/24/facebook-uj-szabalyzat-community-standard>
29. Roy, Ankush Sinha. 2020. “How does Facebook handle the 4+ petabytes of data generated per day? Cambridge Analytica – Facebook data scandal”. *Medium*, September 16. <https://medium.com/@srank2000/how-facebook-handles-the-4-petabyte-of-data-generated-per-day-ab86877956f4>
30. Ruckenstein, Minna and Turunen, Linda Lisa Maria. 2020. „Re-Humanizing the Platform: Content Moderators and the Logic of Care”. *New Media & Society* 6: 1026–1042.
31. Scott, Mark and Kayali, Laura. 2020. „What happened when humans stopped managing social media content”. *Politico*, October 21.
<https://www.politico.eu/article/facebook-content-moderation-automation>
32. Sorbán, Kinga. 2021. „Ethical and legal implications of using AI-powered recommendation systems in streaming services”. *Információs Társadalom* 2: 72–73.
33. Stagnaro, Angelo. 2018. „Facebook Censors Image of Santa Kneeling Before Our Lord” *National Catholic Register*, December 26.
<https://www.ncregister.com/blog/facebook-censors-image-of-santa-kneeling-before-our-lord>
34. Turner Lee, Nicol, Resnick, Paul and Barton, Genie. 2019. „Algorithmic bias detection and mitigation: best practices and policies to reduce consumer harms”. *Brookings Institution*, May 22. <https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms>
35. UNHRC. 2018. *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*. UN Doc A/HRC/38/35, 64.