The Dickson Poon

School of Law

Legal opinion:

The need for due process in Notice-and-takedown-based content moderation on social media platforms

Eva Knoepfel

LLM (General) 2018/2019

Global Digital Enforcement of Intellectual Property (Practice Project), 40 Credit

Supervisor: Prof Frederick Mostert

Word count: 9,567

Abstract

Social media platforms (SMPs) such as Facebook, Twitter and Youtube occupy a uniquely valuable and influential position due to their combined role as the main public forum of global society and the principal regulator of online speech. Although this power comes with a moral and social responsibility to moderate content in such a way that protect users' rights from infringement, Facebook and other major SMPs are routinely criticised for either over-regulating or under-regulating content and thereby directly or indirectly violating users' rights.

Notice-and-takedown (NTD), the primary mechanism used for content moderation, is at the root of many of these problems. However, its deficiencies are exacerbated by SMPs' inconsistent enforcement of their governing rules and insufficiently transparent decision-making. The need to improve content moderation and the protection of users' rights as well as SMPs' role as public platforms and de facto courts for online speech make it both necessary and appropriate for SMPs to adopt due process requirements.

By requiring SMPs to clearly and openly communicate their policy and enforcement decisions; apply their governing rules consistently and without bias; and allow users to present their case and make appeals, due process will improve the protection of users' rights without unnecessarily limiting their freedoms, which is in the interest of users, SMPs and society as a whole.

1.) INTRODUCTION

Content moderators at Facebook, Twitter and Youtube are more powerful than any Supreme Court judge or head of state in deciding whose speech is heard and whose is not.[1] However, the power to regulate online speech is a relentless struggle for social media platforms (SMPs) and more often than not they fail to strike a balance between censoring free speech and creating a lawless cesspool of online abuse.[2]

Notice-and-takedown (NTD), the primary mechanism used for content moderation, is at the root of many of these problems. However, its deficiencies are exacerbated by SMPs' inconsistent enforcement of their governing rules and insufficiently transparent decision-making. The need to improve content moderation and the protection of users' rights as well as SMPs' role as public platforms and de facto courts for online speech makes it both necessary and appropriate for SMPs to adopt due process requirements.

1.1) Modern Roman fora and de facto courts

With almost 3.5 billion users worldwide,[3] SMPs such as Facebook, Twitter and Youtube have become the Roman fora of modern times: hubs of social interaction, information exchange, business and entertainment.[4] Despite their private ownership, these virtual spaces serve such

[1] Jeffrey Rosen, 'The Delete Squad: Google, Twitter, Facebook and the new global battle over the future of free speech' (*The New Republic*, 29 April 2013) <https://newrepublic.com/article/113045/free-speech-internet-silicon-valley-making-rules> accessed 31/08/2019.

[2] Jason Koebler and Joseph Cox, 'The Impossible Job: Inside Facebook's Struggle to Moderate Two Billion People' (*Vice*, 23 August 2018) <https://www.vice.com/en_us/article/xwk9zd/how-facebook-content-moderation-works> accessed 01/09/2019; Jillian C York and Corynne McSherry, 'Content Moderation is Broken. Let Us Count the Ways.' (*EFF*, 29 April 2019) <https://www.eff.org/deeplinks/2019/04/content-moderation-broken-let-us-count-ways> accessed 01/09/2019.

[3] Hootsuite and We Are Social, *Digital 2019 Global Digital Overview* (2019) <https://datareportal.com/reports/digital-2019-global-digital-overview> accessed 31/08/2019.

[4] Frederick Mostert, 'Free Speech and Internet Regulation' (2019) Journal of Intellectual Property Law & Practice, 2.

a wide social and community purpose that they are not merely *a* public platform but *the* public platform of this generation.[5]

To prevent illegal behaviour and create a safe environment for free expression and interaction, SMPs remove content and penalise users for violations of their platform rules and policies, which tend to be based on or reflect the national laws of e.g. the United States (US) or the European Union (EU).[6]

Despite being far from a direct extension of a nation-state's executive or judiciary, SMPs take on the role of a private, informal first-instance court in the regulation of online speech.[7] While users can escalate social media disputes to a national court, there seems to be an understanding that such an "appeal" is only possible once all available remedies at the first instance SMP-level have been exhausted.[8]

On the one hand, SMPs are in a better position to rule on and enforce decisions about user-generated content expeditiously. On the other hand, entrusting Facebook and other major platforms with (near) judicial decision-making powers harbours considerable risks due to the procedures used by SMPs to moderate user content.

Although SMPs increasingly rely on automated AI-based systems, manual reporting and removal of content on the basis of notice-and-takedown (NTD) mechanisms continues to play a significant role in content moderation. However, NTD is an imperfect mechanism for addressing illicit online content, as it lacks procedural guidelines, facilitates fraudulent takedown claims and incentivises internet platforms to over-regulate reported content.[9] The

[5] Ibid.

[6] The term "platform rules" will be used throughout this paper to refer to Facebook's Community Standards, the Twitter Rules and Youtube's Community Guidelines. Platform rules mirror, but are not directly based on any particular jurisdiction's legal offences and thus lack a statutory footing (unlike e.g. DMCA-based copyright claims). When referring to permitted or prohibited behaviour beyond platform rules (e.g. to include copyright) the term "platform rules and policies" will be used; Facebook, *Community Standards* (2019) <https://www.facebook.com/communitystandards/> accessed 09/07/2019.**;** Twitter, *The Twitter Rules* (2019) <https://help.twitter.com/en/rules-and-policies/twitter-rules> accessed 29/08/2019; Youtube, *Community Guidelines* (2019) <https://www.youtube.com/intl/en-GB/yt/about/policies/#community-guidelines> accessed 29/08/2019.

[7] Jack M Balkin, 'Free Speech is a Triangle' (2018) 118(7) Columbia Law Review 2011, 2028-2031.

[8] Ibid 2029; this is of course also a matter of practicality.

[9] Mostert (n4) 8; Frederick Mostert and Jane Lambert, 'Study on IP enforcement measures, especially anti-piracy measures in the digital environment' (WIPO Advisory Committee on Enforcement, 14th Session 2-4 September, Draft paper 10 May 2019) para 15.

combination of NTD and content moderators' power to act as judges compels users to accept a judgement made by a non-judicial authority without any protection against abuse of process.[10] Furthermore, this system relies too much on litigation being an available and viable alternative option when in reality SMPs are often the only arbiters users have access to, and their decisions are therefore not interim but final in effect.[11]

1.2) The need for due process

Due to SMPs' important role as public fora and powerful function as the main (and sometimes only) regulator of the online speech, Facebook and other major platforms have both moral and social obligations towards their users, national governments and society.[12]

When content moderators at Facebook, Twitter or Youtube rule on the legality of user content and make decisions as to its continued existence, they effectively step into the role of judges. Since SMPs act like courts of law which protect and enforce internet users' rights, they ought to follow due process principles[13] to ensure the fair and unbiased administration of justice. While it may not lead to a substantively fair result, procedural justice prevents unreasonable and arbitrary decisions and thus helps increase trust in the decision-making body and the legitimacy of its judgement.[14] By ensuring that justice is not only done but seen to be done,[15] due process increases the likelihood of compliance with a judgement and therefore strengthens the effectiveness of the law or in the case of SMPs, platform rules and policies.[16]

[10] Balkin (n7) 2031.
[11] cf "Convictions cannot be treated as final until appeal rights have been either exhausted or waived", see Peter D Marshall, 'A comparative analysis of the right to appeal' (2011) 22(1) Duke Journal of Comparative & International Law 1.
[12] Balkin (n7) 2041-2045.
[13] This paper uses the terms due process, procedural justice and procedural fairness as interchangeable umbrella terms to refer to the idea that the procedure through which justice is administered must be fair, consistent, independent and transparent. In this sense, the terminology used here reflects the principles of due process as found in the US constitution and natural justice as employed in common law jurisdictions such as the UK.
[14] Rebecca Hollander-Blumoff and Tom R Tyler, 'Procedural Justice and the Rule of Law: Fostering Legitimacy in Alternative Dispute Resolution' (2011) Journal of Dispute Resolution, 3; Tom R Tyler, 'Procedural Justice, Legitimacy, and the Effective Rule of Law' (2003) 30 Crime and Justice 283, 286.
[15] Nina Persak, 'Procedural Justice Elements of Judicial Legitimacy and their Contemporary Challenges' (2016) 6(3) Onati Socio-legal Series 749, 758; Tyler (n14) 286.
[16] Tom R Tyler, *Why People obey the Law* (Princeton University Press 2007) 5-7.

This paper will examine the use of NTD-based content moderation mechanisms by Facebook, Twitter and Youtube and demonstrate that the main obstacle to the effectiveness of this type of online speech regulation is a lack of transparency and the inconsistent enforcement of platform rules and policies.

A theoretical analysis of NTD illustrating the mechanism's inherent flaws (*Chapter 2*) as well as a practical evaluation of different stages in the content moderation process of Facebook, Twitter and Youtube (*Chapter 3*) reveal the need for SMPs to implement due process principles. When NTD is used to moderate context-sensitive content such as copyright infringements and hate speech, procedural justice is crucial to protect users' rights and prevent real-life harm (*Chapter 4*). In particular, SMPs should clearly and openly communicate their policies and enforcement decisions, apply their governing rules consistently and without bias, and offer an option to appeal penalties or request a review of a decision.[17]

2.) NTD IN THEORY

Although NTD aims to strike a fair balance between different rights-holders, it lacks clear procedural guidelines and is designed in such a way that incentivises platforms to over-regulate content, leaves users vulnerable to abusive takedown requests and does not provide for sufficient appeal mechanisms. These systemic flaws can be mitigated by platforms adopting due process standards, which will better protect the rights of platform users.

2.1) From copyright enforcement tool to all-purpose remedy

Originally developed as a tool to remove copyright-infringing content under US law,[18] NTD is now used across the world to address both IP and non-IP law infringements, such as hate speech and defamation.[19] Instead of SMPs and other intermediaries having to monitor user-

[17] These points were identified in the Manila Principles as core principles of due process to be observed by SMPs, see *Manila Principles on Intermediary Liability* (2015) <https://www.eff.org/files/2015/10/31/manila_principles_1.0.pdf> accessed 23/08/2019; Balkin (n7) 2044-2045.

[18] Digital Millennium Copyright Act 17 USC §512, hereafter DMCA.

[19] The Electronic Commerce Directive (ECD) forms the legal basis for NTD procedures in the EU in respect of all types of illicit content, European Parliament and Council, Directive 2000/31/EC on certain legal aspects of

generated content proactively, NTD allows SMPs to avoid liability for infringing material provided they disable access to or remove such content upon being notified of its presence by a purported rights-holder or victim.[20]

However, NTD has come under heavy criticism for facilitating the takedown of legitimate, non-infringing content, thereby undermining freedom of expression and access to information.[21] This "chilling effect" can be attributed to several factors. Firstly, in their effort to discharge liability, intermediaries tend to err on the side of caution and remove reported content too readily.[22] Secondly, NTD gives too much power to rights-holders whose takedown requests may in many cases be based on a mistake (e.g. failure to consider applicable defences such as fair use) or malice (e.g. attempts to silence criticism).[23] Thirdly, internet users whose content is removed are not given adequate opportunity to respond to and appeal takedown notices or the appeal mechanisms available are ineffective.[24] Finally, NTD cannot prevent infringing material from being re-uploaded in a different place (the "whack-a-mole" phenomenon) and therefore offers no a lasting solution to repeat infringements.[25]

2.2) Chilling effect and lack of procedural rules

NTD's chilling effect and other systemic problems are exacerbated by the lack of formal guidance in the underlying legal instruments regarding the application of NTD mechanisms by SMPs. While the DMCA contains some procedural rules and gives alleged infringers a right to appeal the removal of their content by issuing a counter-notification,[26] these provisions do not extend to takedowns on non-copyright grounds which are not covered by the Act. There

information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce'), 8 June 2000; see also Mostert and Lambert (n9) para 11.

[20] The so-called "safe harbour" principle, DMCA §512(c)); see also ECD Arts 14-15.

[21] Christina Angelopoulos and Stijn Smet, 'Notice-and-fair-balance: how to reach a compromise between fundamental rights in European intermediary liability' (2016) 8(2) Journal of Media Law 266, 284-285; Urban *et al.*, *Notice and Takedown in Everyday Practice* (University of California, Berkeley and Columbia University 2017) 2-5.

[22] Mostert (n4) 8.

[23] Mostert and Lambert (n9) para 15.

[24] Urban *et al.* (n21) 44-46.

[25] Mostert and Lambert (n9) para 16.

[26] DMCA §512(g)(3).

is even less clarity on the proper steps to be followed under the ECD, which contains no common EU-wide rules and is silent as to any right to appeal.[27]

However, improved procedural guidance is urgently needed not least in light of repeated calls and concrete plans for NTD to be replaced by a "notice-and-staydown" model to address the "whack-a-mole" phenomenon.[28] The introduction of a system which requires SMPs to filter and block user content is highly controversial and will likely exacerbate, not alleviate, the chilling effect associated with NTD, as SMPs will continue to rely on information provided by rights-holders.[29]

Due process offers a useful and fitting solution to minimise the adverse effects of NTD and create more legal certainty. Not only do SMPs assume the role of judges when moderating content, but the key principles of due process are almost universally recognised as a fundamental tenet of justice[30] and would thus avoid national governments or SMPs having to develop new procedural rules from scratch.[31] If SMPs follow due process and implement their governing rules in a principled and consistent manner and offer users a right to appeal, the risks of content being removed too readily, fraudulent claims and unfair treatment will be significantly reduced.

---

[27] Angelopoulos and Smet (n21) 269; Saskia Walzel, 'European Commission Consults on Notice and Takedown' (*LSE Media Policy Project Blog*, 24 August 2012) <(http://eprints.lse.ac.uk/78705/1/European%20Commission%20Consults%20on%20Notice%20and%20Takedown%20_%20LSE%20Media%20Policy%20Project.pdf> accessed 03/06/2019.

[28] In April 2019, the EU passed the controversial Copyright Directive, Art 17 of which requires SMPs to filter and block new uploads by comparing them against their database of verified, copyrighted content, see European Parliament and Council, *Legislative resolution on the proposal for a directive of the European Parliament and of the Council on copyright in the Digital Single Market*, 26 March 2019 <http://www.europarl.europa.eu/doceo/document/TA-8-2019-0231_EN.html> accessed 09/06/2019; Stephen Carlisle, 'DMCA "Takedown" Notices: Why "Takedown" Should Become "Take Down and Stay Down" and Why It's Good for Everyone' (Nova Southeastern University, 23 July 2014) <http://copyright.nova.edu/dmca-takedown-notices/> accessed 03/05/2019; Mostert and Lambert (n9) para 21.

[29] European Digital Rights, *XSaveYourInternet* (2019) <https://saveyourinternet.eu> accessed 10/06/2019; Mostert and Lambert (n9) para 32.

[30] A succinct summary of the main elements of the principle can be found in the leading Canadian case on procedural fairness in administrative decisions: "The values underlying the duty of procedural fairness relate to the principle that the individual or individuals affected should have the opportunity to present their case fully and fairly, and have decisions affecting their rights, interests, or privileges made using a fair, impartial, and open process, appropriate to the statutory, institutional, and social context of the decision." L'Heureux-Dubé J (p. 841) in *Baker v Canada (Minister of Citizenship and immigration)* [1999] 2 R.C.S., Supreme Court of Canada, No. 25823.

[31] The difficulty of devising such rules is likely the reason for the EU's failed attempts at establishing clearer procedural guidelines, see Mostert and Lambert (n9) para 13.

3.) NTD IN PRACTICE – FACEBOOK, TWITTER AND YOUTUBE[32]

Facebook, Twitter and Youtube are three of the oldest and most frequented SMPs.[33] Furthermore, all three have all been accused of both over-regulating speech and taking down legitimate content,[34] and under-regulating content and creating a fertile breeding ground for abuse.[35] The reasons for these failures in content moderation are mainly procedural in nature: a lack of transparency regarding the enforcement of platform rules and policies, the use of AI, the review stage and penalties; uneven enforcement of rules and policies and unequal treatment of users; and insufficient appeal mechanisms.

Many, if not all, of these deficiencies can be remedied by implementing due process requirements at each stage of the content moderation process.

3.1) Platform rules and policies and transparency reports

Platform rules and policies are the applicable law on Facebook, Twitter and Youtube and must thus be clear, precise and easily accessible to users to be effective and complied with.[36]

In response to demands for more transparency by governments, NGOs and initiatives such as the Santa Clara Principles,[37] Facebook, Twitter and Youtube have not only published their platform rules[38] but also released transparency reports including numbers and statistics on enforcement.[39]

---

[32] The rules and policies analysed in this and the following chapter were in effect at the time of writing and, to the best of the author's knowledge, at the time of submission of this paper.

[33] Hootsuite and We Are Social (n3).

[34] see e.g. Sam Wolfson, 'Facebook labels declaration of independence as 'hate speech'' *The Guardian* (London, 5 July 2018) <https://www.theguardian.com/world/2018/jul/05/facebook-declaration-of-independence-hate-speech> accessed 10/07/2019.

[35] see e.g. Alex Warofka (Facebook Product Policy Manager), 'An Independent Assessment of the Human Rights Impact of Facebook in Myanmar' (*Facebook Newsroom*, 5 November 2018) <https://newsroom.fb.com/news/2018/11/myanmar-hria/> accessed 10/07/2019.

[36] James R Maxeiner, 'Some Realism about Legal Certainty in the Globalization of the Rule of Law' (2008) 31(1) Houston Journal of International Law 27.

[37] The Santa Clara Principles are the result of a joint effort of civil society organisations, advocates and academics to put forward minimum standards of transparency, accountability and due process which internet platforms should observe, see **Appendix** or *The Santa Clara Principles On Transparency and Accountability in Content Moderation* (2018) <https://santaclaraprinciples.org> accessed 24/08/2019.

[38] Facebook (n6); Twitter (n6); Youtube (n6).

[39] Facebook, *Facebook Transparency Report* (2019) <https://transparency.facebook.com> accessed 01/09/2019; Google, *YouTube Community Guidelines enforcement* (2019)

Facebook, Twitter and Youtube's rules show considerable overlap in terms of prohibited content and behaviour, which includes spam, hate speech, violent threats and the promotion of terrorism (platform rules), IP law infringements (copyright, trademark, counterfeits) and criminal offences such as child pornography.[40] The only notable difference concerns graphic violence and pornography/adult nudity, which is banned on Facebook and Youtube but permitted on Twitter in certain circumstances.[41]

However, contrary to one of the three key demands of the Santa Clara principles,[42] the platforms do not report their numbers in a uniform manner or use a common denominator which allows for a comparison of their statistics. While Twitter's platform rules report refers only to the number of user accounts affected, Facebook and Youtube only report the number of pieces of content actioned.[43]

Similar issues affect the reports on copyright takedown requests and removals. While Facebook reports the number of reported infringements, actual takedowns and removal rate, Twitter also presents DMCA counter-notices in its report.[44] Youtube, on the other hand, does not include any data on IP-law based removals in its transparency report, which is a severe shortcoming considering the amount of criticism the platform receives over its copyright policies.[45]

<https://transparencyreport.google.com/youtube-policy/removals?hl=en> accessed 09/07/2019; Twitter, *Transparency Report* (2019) <https://transparency.twitter.com> accessed 01/09/2019.

[40] Facebook, *Facebook Terms and Policies* (2019) <https://en-gb.facebook.com/policies> accessed 01/09/2019; Twitter, *Twitter Rules and policies* (2019) <https://help.twitter.com/en/rules-and-policies#general-policies> accessed 01/09/2019; Youtube, *Youtube Policies* (2019) <https://support.google.com/youtube/topic/2803176?hl=en-GB&ref_topic=6151248,3230811,3256124,> accessed 01/09/2019.

[41] Twitter, *Sensitive media policy* (2019) <https://help.twitter.com/en/rules-and-policies/media-policy> accessed 01/07/2019.

[42] Numbers, notice and appeal, see The Santa Clara Principles (n37).

[43] Facebook, 'Community Standards Enforcement Report' *Facebook Transparency Report* (2019) <https://transparency.facebook.com/community-standards-enforcement> accessed 01/07/2019; Google (n39); Twitter, 'Twitter Rules enforcement' *Transparency Report* (2019) <https://transparency.twitter.com/en/twitter-rules-enforcement.html> accessed 24/08/2019.

[44] Facebook, 'Intellectual Property' *Facebook Transparency report* (2019) <https://transparency.facebook.com/intellectual-property> accessed 24/08/2019; Twitter, 'Copyright notices' *Transparency report* (2019) <https://transparency.twitter.com/en/copyright-notices.html> accessed 23/08/2019.

[45] For a more detailed discussion, see **4.1) Youtube and Copyright.**

Finally, none of the platforms has so far published both the number of platform rule appeals received and the number of copyright counter-notifications, which is why a cross-platform assessment of the effectiveness of appeal mechanisms for IP and non-IP law violations is not possible.

3.2) Detecting and reporting violations

Facebook, Youtube and Twitter use a combination of AI-based machine detection and human flagging to detect and remove infringing content. The platforms' transparency reports suggest that AI is widely used to identify platform rule violations[46] and has a high proactive detection rate in policy areas which are not context-sensitive.[47] However, none of the platforms provides any details on the nature and extent of their AI-use. This lack of concrete information about where, when, and how much AI is used is a considerable drawback when it comes to identifying the reasons for wrongful takedowns and the solutions required to prevent a reoccurrence. A post which was mistakenly removed because of an algorithm's inability to understand sarcasm requires a different response than a post which was deliberately wrongfully flagged as a violation by another user.

The impact of automated tools for the detection of copyright infringements such as Youtube's Content ID system is similarly opaque.[48] Although these mechanisms are only available to a small number of eligible users, they include large film and record companies who own the rights to a substantial body of work.[49] With a large number of copyrighted works automatically identifiable, more alleged violations can be detected than through manual reports. Therefore, the exclusion of Content ID and Rights Manager claims from Youtube and

[46] e.g. Youtube reports that the vast majority of its removed videos (approx. 70-80%) and comments (approx. 99%) were detected by its AI flagging system, see Google (n39), 'Removed videos by the numbers, October 2017-March 2019' and 'Removed comments by the numbers, July 2018-March 2019'.
[47] e.g. Facebook's proactive detection rate for spam has consistently been at 99% over the past couple of years, compared to only 14-21% for bullying and harassment: see Facebook (n43).
[48] Content ID, Youtube's digital fingerprinting system, is one a number of automated tools to facilitate the detection and management of copyrighted works, see Youtube, *How Content ID works* (2019) <https://support.google.com/youtube/answer/2797370> accessed 24/08/2019; Facebook's version of Content ID is its Rights manager, see Facebook, *Rights Manager* (2019) <https://rightsmanager.fb.com> accessed 01/09/2019.
[49] Ibid.

Facebook's transparency reports creates an incomplete and potentially distorted picture of the prevalence of copyright infringements on the platforms and of automated systems' effectiveness in identifying them.

In contrast to the lack of transparency surrounding the use of automated systems, Facebook, Twitter and Youtube publish the numbers of reported and actioned violations of their platform rules as well as copyright takedown notices and counter-notices based on the DMCA. The reports show that platform rule violations manually reported or "flagged" by users have a high error rate on all three platforms regardless of the reporting reason selected, which demonstrates the importance of review before any action is taken.[50]

In respect of copyright, Youtube (unlike Twitter and Facebook) does not state that its takedown request form is based on the DMCA despite using the same wording as the Act.[51] Since a claim under the DMCA has serious legal consequences for the reporting party,[52] it should not be lightly assumed that Youtube users around the world will recognise the claim form's basis in US copyright law and understand its implications.

3.3) Review stage

The review of reported violations is the most crucial stage in the content moderation process for it results in a "judgement" on whether a violation has indeed occurred. To ensure that justice is not only done but seen to be done, this critical decision-making process needs to be transparent and follow clearly defined procedural rules.

The general overarching review framework is openly and clearly communicated to users at Facebook, Twitter and Youtube. Before any action is taken, user-reported violations are manually reviewed[53] as to formal accuracy and completeness (DMCA takedown requests) or

[50] Facebook (n43); Google (n39); Twitter (n43).
[51] Youtube, *Copyright Infringement Notification* (2019) <https://www.youtube.com/copyright_complaint_form> accessed 24/08/2019.
[52] See **4.1) Youtube and copyright.**
[53] Facebook claims that it manually reviews *all* flagged content (AI and human) before taking any action, Facebook, *Understanding the Community Standards Enforcement Report* (2019) <https://transparency.facebook.com/community-standards-enforcement/guide#section4> accessed

substance (platform rule violations). When doing so, the content is evaluated against all platform rules irrespective of the original flagging reason, and content moderators must take into account contextual elements such as intent.[54]

By contrast, the three platforms have been very reluctant to disclose the underlying mechanisms of the review process, which in turn has contributed to many users' perception that decisions about content are made in an arbitrary, haphazard manner. However, this negative impression may be largely unjustified, at least as far as Facebook is concerned. As various investigative reports, leaked documents and former employees have revealed, Facebook employs thousands of content moderators across the globe to review every piece of reported content on the basis of detailed internal guidelines which are constantly updated to reflect current events and developments.[55]

As has been pointed out by Professor Kate Klonick, this review process requires the "*exercise [of] professional judgment concerning the application of a platform's internal rules and… [the] use [of] legal concepts like relevance, reason through example and analogy*".[56] In this sense, the role of content moderators is very reminiscent of that of a judge.[57] However, unlike judges, Facebook's content reviewers' identity and work are deliberately kept hidden from the public, if necessary by using NDAs.[58]

01/09/2019. By contrast, the approach taken by Twitter and Youtube is less clear, although it appears that at least most, if not all, human-flagged content is reviewed, whereas many types of machine-flagged items are removed automatically, such as spam, see TeamYouTube[Help], 'The Life of a Flag' (*Youtube*, 23 April 2018) <https://www.youtube.com/watch?v=WK8qRNSmhEU> accessed 11/07/2019.

[54] Further factors may include the identity of the reporter (victim or bystander) and target (individual, group, protected characteristic), see Youtube, *The Importance of Context* (2019) <https://support.google.com/youtube/answer/6345162?hl=en> accessed 17/07/2019 ; Google, 'Human flags by flagging reasons' *Youtube Community Guidelines enforcement report* (2019) <https://transparencyreport.google.com/youtube-policy/flags> accessed 23/08/2019; Twitter, *Our approach to policy development and enforcement philosophy* (2019) <https://help.twitter.com/en/rules-and-policies/enforcement-philosophy> accessed 17/07/2019.

[55] see e.g. Adrian Chen, 'The laborers who keep dick pics and beheadings out of your Facebook feed' (*Wired*, 23 October 2014) <https://www.wired.com/2014/10/content-moderation/> accessed 17/07/2019; Casey Newton, 'The Trauma Floor - The secret lives of Facebook moderators in America' (*The Verge*, 25 February 2019) <https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona> accessed 16/06/2019.

[56] Kate Klonick, 'The new governors: the people, rules, and processes governing online speech' (2018) 131 Harvard Law Review 1598, 1641-1642.

[57] Ibid.

[58] Newton (n55).

While some level of discretion may be necessary for data privacy purposes and the personal safety of moderators, Facebook and other major platforms' secretiveness about their internal processes

also shields them from criticism and liability, which many users have now begun to call out.[59] Transparency in content review is not merely for the benefit of platform users who demand to know how decisions about their rights online are made. It is also in the interest of SMPs that the difficult but essential work of content moderators is seen and acknowledged, particularly when decisions about controversial material have to be made. Since content moderation inevitably involves a balancing of competing rights, the substantive outcome will often be unsatisfactory for at least one party. It is therefore crucial that the platforms ensure that at least the procedure through which an outcome is reached is as open and transparent as possible.

3.4) Enforcement and penalties

Once a violation has been confirmed or refuted, the material in question may be left up, restricted or removed and the responsible user will usually face a penalty for breaching the platform's rules and policies.

Although the most appropriate action to take will depend on the type of breach and enforcement options available to each platform, the sanctions imposed ought to be reasonably predictable and proportionate to the violation committed. Additionally, all users should be treated equally and any exceptions made need to be justified and kept to a minimum.

Youtube's community guidelines and copyright three-strikes systems clearly outline the penalties that each strike will attract but have been criticised for being too harsh and prone to abuse.[60] Particularly the copyright strike system in particular has proven to be an easy

[59] Ibid.

[60] Community guideline strikes lead to a temporary suspension of the user's privileges (e.g. uploading videos), whereas copyright strikes affect an account's good standing and may disable the monetisation feature, see Youtube, *Strikes FAQ* (2019) <https://support.google.com/youtube/answer/9235777?hl=en&ref_topic=2803138> accessed 23/08/2019; Youtube, *Community Guideline strikes basics* (2019) <https://support.google.com/youtube/answer/2802032?hl=en> accessed 23/08/2019.

method to silence or harm content creators since the merits of DMCA takedown notices which give rise to a copyright strike do not need to be examined by Youtube.[61]

In contrast to Youtube, Facebook neither reveals the precise sanctions that users will face for receiving a strike nor how many strikes will cause an account to be permanently suspended, arguing that "[w]e don't want people to game the system".[62] Not knowing the nature or severity of one's punishment may act as a deterrent in certain circumstances but might also create the impression of an arbitrary system of justice which is entirely at the discretion of Facebook's content moderation team.

While Twitter and Facebook have come under criticism for not taking action on violating content because of its "newsworthiness",[63] Youtube has been accused of treating users with a public profile more favourably when enforcing its platform rules.[64]

Although Twitter claims that the newsworthiness exception is very rarely used, it seems to benefit exclusively public figures with a large following, such as US President Donald Trump. Twitter's plans to label Tweets which breach its rules but are left live for public interest purposes is likely to create more transparency and signal to other users that such content or behaviour is not ordinarily tolerated.[65] At the same time, such a measure would undoubtedly also act as a visual reminder that Twitter does not treat all of its users equally.

---

[61] Examples of particularly brazen "bogus" DMCA claims can be found on EFF's Takedown Hall of Shame: EFF, *Takedown Hall of Shame* (2019) <https://www.eff.org/takedowns/> accessed 01/09/2019. Most recently, one individual issued so many fraudulent DMCA claims and threatened Youtubers with copyright strikes unless they paid him a ransom, that Youtube issued a lawsuit against him, see Youtube, *Youtube v Christopher L Brady: Demand for Jury Trial* (United States District Court District of Nebraska, Case No. 19-353, 19 August 2019) <https://torrentfreak.com/images/Youtube-v-Christopher-Brady-DMCA-abuse-complaint-191908.pdf> accessed 01/09/2019; Katharine Trendacosta, 'YouTube's New Lawsuit Shows Just How Far Copyright Trolls Have to Go Before They're Stopped' (*EFF*, 21 August 2019) <https://www.eff.org/deeplinks/2019/08/youtubes-new-lawsuit-shows-just-how-far-copyright-trolls-have-go-theyre-stopped> accessed 01/09/2019.

[62] Facebook, 'Enforcing Our Community Standards' (*Facebook Newsroom,* 6 August 2018) <https://newsroom.fb.com/news/2018/08/enforcing-our-community-standards/> accessed 17/07/2019.

[63] Facebook (n6); Twitter (n54).

[64] Elizabeth Dwoskin, 'YouTube's arbitrary standards: Stars keep making money even after breaking the rules' *The Washington Post* (Washington DC, 9 August 2019) <https://www.washingtonpost.com/technology/2019/08/09/youtubes-arbitrary-standards-stars-keep-making-money-even-after-breaking-rules/> accessed 01/09/2019.

[65] Isobel Asher Hamilton, 'Twitter wants to label tweets from public figures that break its rules — and even Trump could be named and shamed' (*Business Insider*, 29 March 2019) <https://www.businessinsider.com/twitter-to-label-tweets-from-public-figures-like-trump-that-violate-rules-2019-3?r=US&IR=T> accessed 23/08/2019.

3.5) Appeals

The nature of NTD requires SMPs to action allegedly infringing content without first offering the responsible user an opportunity to defend themselves. To minimise the risk of injustice, platforms should protect users' right to be heard by offering an easy and effective way to appeal and reverse content moderation decisions.

Facebook, Twitter and Youtube state that they inform each user who has breached their rules or policies of the specific provision violated, the penalty imposed and how the decision can be appealed, which in theory should make the appeal process relatively straight-forward.[66] This is indeed the case for copyright claims which can be appealed on all three platforms by submitting a DMCA counter-notification.[67]

Like the number of DMCA takedown notices, the number of counter-notices submitted on Twitter has been steadily increasing over the years, from 16 in 2013, to 516 in 2016 to 2,214 in 2018.[68] Furthermore, in all but one instance, the submission of a counter-notice has led to the reinstatement of the content in question.[69] However, this high restoration rate may not speak as much to the effectiveness of counter-notices as it does to the difficulty of initiating court proceedings within ten business days to prevent such reinstatement.[70]

By comparison, the appeals process for platform rule violations on Facebook has been described by users as exceedingly cumbersome. According to these reports, the option to appeal a decision is frequently not available when users are notified of their violation, and the free-standing appeals form on Facebook is very difficult to locate, as is any information on Facebook's appeal policy.[71]

---

[66] Facebook (n6); Twitter (n43); Youtube (n60).

[67] Facebook, *Appealing a Claim of Copyright Infringement Made Under the DMCA (Counter-Notification)* (2019) <https://www.facebook.com/legal/copyright.php?howto_appeal=1> accessed 01/09/2019; Twitter, 'Copyright policy' *Twitter Rules and policies* (2019) <https://help.twitter.com/en/rules-and-policies/copyright-policy> accessed 01/09/2019; Youtube, *Copyright counter notification basics* (2019) <https://support.google.com/youtube/answer/2807684?hl=en-GB&ref_topic=9282678> accessed 01/09/2019.

[68] Twitter (n44).

[69] Ibid.

[70] DMCA §512(g)(2)(c).

[71] see Facebook users' repeated requests for a direct link to an appeals form in the Facebook forum, see *Facebook Community forum*, 'Heather's question' (10 October 2018) <https://m.facebook.com/help/community/question/?id=158432015100013&rdrhc> accessed 18/07/2019; 'Julie's question' (16 August 2018) <

Twitter is the only platform which allows users to explain why they believe their content did not breach the Twitter Rules.[72] Furthermore, by making it possible for users to submit appeals directly from within the Twitter app, the platform is able to respond to appeals 60% faster than before when appeals could only be made through an online form.[73]

While Facebook, Twitter and Youtube have taken some important steps towards adopting more procedurally fair processes, the overall NTD-based content moderation system used by the three platforms is still a long way away from being procedurally just.

4.) USING A FLAWED SYSTEM TO MODERATE COMPLEX CONTENT: COPYRIGHT INFRINGEMENTS AND HATE SPEECH

The procedurally flawed NTD systems of Facebook, Twitter and Youtube pose additional challenges when it comes to moderating complex, context-sensitive content and call for more transparent and consistent enforcement of the platforms' rules and policies. In the case of copyright infringements on Youtube, the need for more procedural fairness arises because the platform is often the first and only arbiter users have access to given the difficulty of bringing a copyright case to court. By contrast, due process is required in the moderation of hate speech to minimise the risk of real-life violence, which is in the best interests of users, the platforms and wider society.

https://m.facebook.com/help/community/question/?id=10215596085529558&rdrhc> accessed 18/07/2019; 'Francie's question' (17 August 2018) < https://m.facebook.com/help/community/question/?id=1726991954016707&answer_id=1744943225554913 > accessed 18/07/2019.

[72] The example used by Twitter is that of a user explaining that their Tweet was referring to a joke about a video game and was thus not an actual threat which violates the rules against abusive behaviour, see Twitter Safety, 'Introduction of in-app appeal mechanism' (*Twitter*, 2 April 2019) <https://twitter.com/TwitterSafety/status/1113139073303089152> accessed 11/07/2019.

[73] Ibid.

4.1) Youtube and copyright

Even though copyright disputes have plagued Youtube since its inception, only a handful of cases have made it to court.[74] This lack of litigation is indicative of the obstacles users face in taking such legal action and which may compromise their access to justice in national courts. To prevent users from being denied a remedy, it is vital that Youtube exercises its role as a de-facto court with due process.

4.1.i) The DMCA-based copyright takedown mechanism

The fact that Youtube's NTD system for copyright is based on the DMCA puts non-US users at a considerable disadvantage since any counter-notification issued under the Act requires the alleged infringer to consent to the jurisdiction of a US federal district court. As a result of this provision, copyright holders as well as alleged infringers put themselves at risk of having to litigation in the US by using Youtube's copyright takedown or counter-notice form, and incur considerable expenses and loss of time.[75]

Given these stakes, Youtube's failure to state the legal basis of its takedown notice form constitutes a serious lack of transparency, which may result in an unpleasant surprise for unsuspecting users.[76] By contrast, users who are familiar with the DMCA and its jurisdictional implications may be forced to forego the chance to protect their copyrighted works or resist unsubstantiated claims.

4.1.ii) The complex nature of "fair use"

The lack of litigation in Youtube copyright disputes may also be attributable to the complex nature of the "fair use" defence under US copyright law which permits the limited use of

[74] See e.g. Noam Cohen, 'YouTube Is Purging Copyrighted Clips' *The New York Times* (New York, 30 October 2006) <https://www.nytimes.com/2006/10/30/technology/30youtube.html> accessed 19/08/2019; Jonathan Bailey, 'YouTube's Copyright Problem' (*Plagiarism Today*, 23 October 2013) <https://www.plagiarismtoday.com/2013/10/23/youtubes-copyright-problem/> accessed 19/08/2019; Jake Plovanic, 'YouTube (Still) Has a Copyright Problem' Washington Journal of Law, Technology & Arts (*WJLTA Blog*, 28 February 2019) <https://wjlta.com/2019/02/28/youtube-still-has-a-copyright-problem/> accessed 19/08/2019.

[75] Keli Johnson Swan, 'United States: The True Cost Of Defending Against Copyright Infringement Litigation' (*Mondaq*, 19 August 2015) <http://www.mondaq.com/unitedstates/x/421188/Copyright/The+True+Cost+Of+Defending+Against+Copyright+Infringement+Litigation> accessed 01/09/2019.

[76] Youtube (n51).

copyrighted works without prior permission from the copyright owner.[77] However, a finding of fair use necessitates a careful balancing of the rights of the copyright owner against those of the secondary user based on multiple factors which may be weighed differently depending on the context of each case.[78] The flexibility of fair use comes at the cost of legal certainty due to the absence of fixed requirements.[79] Furthermore, there is hardly any precedent for Youtube creators to rely on and the few cases which have been decided may be distinguishable on the facts.[80] As a result, it is very difficult for users to predict whether their video qualifies as fair use and any disputes taken to court will likely result in lengthy proceedings and high legal fees which many users cannot afford.[81] Particularly in cases which "only" concern the removal of a single video, the cost of litigation is grossly disproportionate to any potential gains.[82]

4.1.iii) Unequal opponents

The fact that DMCA takedown notices and Content ID claims are often issued by global entertainment companies who have both the financial resources and manpower to pursue litigation forces many defendants to concede early defeat even if their claim has merit.[83] Music record companies appear to be well-aware of and frequently use their superior bargaining position to demand the removal or claim the revenue of videos which contain only tiny portions of copyrighted material or which might be covered by fair use.[84]

[77] Copyright Act of 1976 17 USC §107.
[78] Ibid; The four main factors to be considered are (1) the purpose and character of the use, (2) the nature of the original/copyrighted work, (3) the amount and substantiality of the portion of the original work used and (4) the effect of this use on the original work.
[79] Elliot Harmon, 'Don't Sacrifice Fair Use to the Bots' (*EFF*, 1 March 2019) <https://www.eff.org/deeplinks/2019/03/dont-sacrifice-fair-use-bots> accessed 19/08/2019.
[80] *Hosseinzadeh v Klein*, 276 F. Supp. 3d 34 (S.D.N.Y. 2017) (hereafter the "h3h3 case") at 41, Footnote 1.
[81] Youtubers Ethan and Hila Klein, the claimants in the h3h3 case, had to rely on crowd-funding to pay for their attorney fees, which at one point rose to over $50,000 per month, see h3h3Productions, 'We're Still Being Sued' (27 February 2017) <https://www.youtube.com/watch?v=m40bWgWH8Ro> accessed 20/08/2019.
[82] Taylor B Bartholomew, 'The Death of Fair Use in Cyberspace: Youtube and the Problem with Content ID' 13(1) Duke Law & Technology Review 66, 85.
[83] Julia Alexander, 'Youtubers and record lables are fighting, and record labels keep winning' (*The Verge*, 24 May 2019) <https://www.theverge.com/2019/5/24/18635904/copyright-youtube-creators-dmca-takedown-fair-use-music-cover> accessed 19/08/2019.
[84] In response to this type of aggressive claiming by record companies, Youtube recently changed its policy so as to no longer allow copyright owners to claim the advertising revenue of videos which feature only a short extract of copyrighted music. However, the copyright owner can still demand that such videos be taken down, see Youtube, 'Updates to our manual Content ID claiming policies' (*Youtube Creator Blog*, 15 August 2019) <https://youtube-creators.googleblog.com/2019/08/updates-to-manual-claiming-policies.html> accessed 01/09/2019.

Even though copyright holders are required to consider fair use before issuing takedown notices, they need only have a good faith belief that an infringement has occurred.[85] A lack of this highly subjective good faith requirement is very difficult and therefore costly and time-consuming for a defendant to prove.[86] Thus far, it appears that only Youtubers with a large audience have successfully managed to change a record company's subjective belief that their video was copyright-infringing.[87] The fact that copyright owners and/or Youtube will only reconsider the removal or monetisation of a video if the video creator in question has a large enough fan base demonstrates that not all Youtube users are treated equally.

Although any action taken by Youtube in response to a copyright takedown request or counter-notice is in theory only an interim decision, in practice it will amount to a final verdict if the option of appealing the decision in court is not accessible.[88]

If Youtube is the only arbiter or judge that users realistically have access to, it is crucial that the platform makes procedurally just decisions, adheres to clearly defined and transparent procedures and delivers results which users can easily comprehend.

4.2) Facebook and hate speech

The correlation between hate speech and real-life violence has been well observed throughout history.[89] On social media, algorithms and filter bubbles tend to amplify the risk of virtual hate speech inciting physical violence, which is why SMPs must swiftly and

[85] *Lenz v Universal Music Corp*, 801 F.3d 1126 (2015), (9th Cir. 2015), also known as the "dancing baby" case.
[86] Ibid; *Lenz v Universal Music Corp* was litigated for over 10 years until the parties reached a settlement, see Corynne McSherry, 'After More Than a Decade of Litigation, the Dancing Baby Has Done His Part to Strengthen Fair Use for Everyone' (*EFF*, 27 June 2018) <https://www.eff.org/deeplinks/2018/06/after-more-decade-litigation-dancing-baby-ready-move> accessed 19/08/2019.
[87] For instance, after Youtube creator James Charles complained on Twitter to his millions of followers about receiving a Content ID claim, the music production company in question decided to release its claim, see Benedict Townsend, 'James Charles and his fans win copyright battle against YouTube' (*We The Unicorns*, 9 April 2019) <https://www.wetheunicorns.com/youtubers/james-charles/james-charles-fans-win-copyright-battle/> accessed 19/08/2019.
[88] Marshall (n11).
[89] For instance, the Nazis' use of anti-Semitic propaganda to garner support for the Holocaust or the "radio hate campaign" in the 1994 Rwandan genocide calling for the extermination of the Tutsi minority, see e.g. Adena *et al.*, 'Radio and the Rise of The Nazis in Prewar Germany' (2015) 130(4) The Quarterly Journal of Economics 1885; David Yanagizawa-Drott, 'Propaganda and conflict: evidence from the Rwandan Genocide' (2014) 129(4) Quarterly Journal of Economics 1947.

consistently remove content which violates their rules against hate speech.[90] As the case of the Rohingya people in Myanmar demonstrates, Facebook's inaction in the face rampant hate speech, non-transparent decision-making and ill-conceived responses can have dramatic consequences and contribute to "textbook ethnic cleansing".[91]

4.2.i) Underestimating the situation in Myanmar and Facebook's role

Notwithstanding the complexity of Myanmar's decade-long internal conflicts, once reports about human rights violations and their connection to hate speech on Facebook started surfacing, the platform should have been on the alert and begun to put safeguards in place.[92] At the very least, Facebook ought to have immediately increased the number of Burmese speaking content moderators, which has been pitifully low in the past,[93] considering Facebook' immense reach and popularity in Myanmar.[94]

Facebook's decisions as to which political groups to ban were seen as controversial, however, as the independent human rights assessment report commissioned by Facebook in late 2018 revealed, how these decisions were reached and subsequently enforced were considered even more problematic.[95] In particular, Facebook failed to involve local stakeholders who had

[90] A recent study discovered a direct causal link between anti-immigration hate speech on Facebook and attacks on refugees in Germany, see Karsten Müller and Carlo Schwarz, *Fanning the Flames of Hate: Social Media and Hate Crime* (University of Warwick CAGE Working Paper Series No.373, May 2018) <https://warwick.ac.uk/fac/soc/economics/research/centres/cage/manage/publications/373-2018_schwarz.pdf> accessed 21/08/2019.

[91] Michael Safi, 'Myanmar treatment of Rohingya looks like 'textbook ethnic cleansing', says UN' *The Guardian* (London, 11 September 2017) <https://www.theguardian.com/world/2017/sep/11/un-myanmars-treatment-of-rohingya-textbook-example-of-ethnic-cleansing> accessed 22/08/2019.

[92] Myanmar Civil Society Organizations, *Open Letter to Mark Zuckerberg* (5 April 2018) <https://drive.google.com/file/d/1Rs02G96Y9w5dpX0Vf1LjWp6B9mp32VY-/view> accessed 24/08/2019; Maya Kosoff, 'Facebook's Hate-speech problem may be bigger than it realized' (*Vanity Fair,* 21 August 2018) <https://www.vanityfair.com/news/2018/08/facebooks-hate-speech-problem-may-be-bigger-than-it-realized> accessed 24/08/2019.

[93] Until at least 2015 Facebook only employed two Burmese-speaking moderators, see Steve Stecklow, *Why Facebook is losing the war on hate speech in Myanmar* (Reuters Special report, 15 August 2018) <https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/> accessed 24/08/2019.

[94] It is currently estimated that 20 million of Myanmar's 53 million citizens use Facebook with many considering the platform to be the internet, see Elise Thomas, 'Facebook Keeps Failing in Myanmar' (*Foreign Policy*, 21 June 2019) <https://foreignpolicy.com/2019/06/21/facebook-keeps-failing-in-myanmar-zuckerberg-arakan-army-rakhine/> accessed 20/08/2019; UN Human Rights Council, *Report of the independent international fact-finding mission on Myanmar* (A/HRC/39/64, 12 September 2018), para 74 <https://www.ohchr.org/Documents/HRBodies/HRCouncil/FFM-Myanmar/A_HRC_39_64.pdf> accessed 24/08/2019.

[95] BSR, *Human Rights Impact Assessment: Facebook in Myanmar* (October 2018) <https://fbnewsroomus.files.wordpress.com/2018/11/bsr-facebook-myanmar-hria_final.pdf> accessed 24/08/2019.

a better contextual understanding of the conflict in the development of its policies.[96] Additionally, while most stakeholders deemed Facebook's existing rules against hate speech appropriate, they perceived their inconsistent enforcement as a significant barrier to successfully addressing the hateful content against the Rohingya.[97]

Finally, the continued banning of groups and individuals without prior consultation of civil society groups and the poor enforcement of these new policies suggest a lack of commitment by Facebook to its own policy decisions and an attempt to evade responsibility by avoiding their enforcement. [98]

4.2.ii) Keeping out of politics

After being praised for facilitating political reform in the Arab spring, Facebook has been facing growing pressure to stay out of political affairs. In addition to the public and political backlash over its contribution to the Cambridge Analytica Scandal,[99] Facebook has faced repeated accusations of having a liberal bias,[100] all of which may have contributed to its delayed response to the events in Myanmar.

While a detailed discussion of Facebook's role in politics is beyond the remit of this paper, it is essential to consider that regardless of Facebook's efforts to reduce bias in the implementation of its community standards, it is unlikely ever to be perceived as a neutral actor when moderating polarising political issues. For this reason, Facebook must be open and transparent about the interests and values driving its decisions as well as the process and reasoning behind its policies to avoid creating the impression of bias, ulterior motives and lack of legitimacy.[101]

---

[96] Ibid 26.

[97] Ibid.

[98] Thomas (n94); Julia Carrie Wong, ''Overreacting to failure': Facebook's new Myanmar strategy baffles local activists' *The Guardian* (London, 7 February 2019) <https://www.theguardian.com/technology/2019/feb/07/facebook-myanmar-genocide-violence-hate-speech> accessed 24/08/2019.

[99] see e.g. Mark Zuckerberg, *Witness Testimony* (Facebook, Social Media Privacy, and the Use and Abuse of Data: Hearing before the United States Senate Committee on the Judiciary and the United States Senate Committee on Commerce, Science and Transportation, 10 April 2018) <https://www.judiciary.senate.gov/imo/media/doc/04-10-18%20Zuckerberg%20Testimony.pdf> accessed 24/08/2019.

[100] Although an independent audit into Facebook's alleged anti-conservative bias did not find evidence of systemic political bias, it emphasised the need for more transparency in Facebook's content moderation process: Jon Kyl, *Covington Interim Report* (2019) <https://fbnewsroomus.files.wordpress.com/2019/08/covington-interim-report-1.pdf> accessed 24/08/2019.

[101] Wong (n98).

4.2.iii) The extent of Facebook's fault

Unlike radios which helped hate speech spread in the Rwandan genocide, Facebook does not merely transmit but curate speech by determining through content moderation and algorithms what content users are exposed to.[102] Much of the anti-Rohingya propaganda fell within Facebook's definition of hate speech and had Facebook enforced its platform rules consistently, much of the offending content would have been swiftly removed and further harm could have been avoided.[103]

At the same time, however, Facebook should not be made a scapegoat for the atrocities committed against the Rohingya people, nor should its role be overstated to distract from the government's failure to protects its citizens, put a stop to the violence and bring those responsible to justice.[104]

While Facebook may not be held formally accountable for its failures in Myanmar, it is highly likely that it will face stricter regulation and penalties if it does not drastically improve the enforcement of its hate speech rules.[105] Above all, Facebook needs to pursue a more principled approach to dealing with harmful content such as hate speech; one that is transparent to the public, open to critique and avoids the impression that Facebook is complicit in conflict, biased towards a party or simply does not care. The adoption of due process standards in content moderation would therefore be in the interest of both Facebook and of its users.

---

[102] Balkin (n7).

[103] Stecklow (n93).

[104] Milton Mueller, *Challenging the Social Media Moral Panic* (Cato Institute Policy Analysis No. 876, 23 July 2019) <https://object.cato.org/sites/cato.org/files/pubs/pdf/pa-876-update.pdf> accessed 21/08/2019; John Samples, 'False Assumptions Behind the Current Drive to Regulate Social Media' (*Cato Institute Blog*, 23 July 2019) <https://www.cato.org/blog/false-assumptions-behind-current-drive-regulate-social-media> accessed 21/08/2019).

[105] France recently voted in favour of introducing an online hate speech law which requires SMPs to take down "obviously hateful" content within 24 hours or face a fine of up to €1.25m. Facebook was ordered to pay a €2m fine for breaching a similar law in Germany (NetzDG) which has been in force since 2018, see James McAuley**,** 'France moves toward a law requiring Facebook to delete hate speech within 24 hours' *The Washington Post* (Washington DC, 9 July 2019) <https://www.washingtonpost.com/world/europe/france-moves-toward-a-law-requring-facebook-to-delete-hate-speech-within-24-hours/2019/07/09/d43b24c2-a25d-11e9-a767-d7ab84aef3e9_story.html> accessed 24/08/2019.

5.) SUGGESTIONS AND CONCLUSION

5.1) Summary of findings

The unique role of SMPs as the main public fora of modern society on the one hand, and the primary regulators of online speech on the other hand, places an obligation these platforms to ensure that users are able to exercise their rights without infringing on the freedoms of others. However, as this paper has demonstrated, this is a task that SMPs often fail at. On the one hand, the NTD mechanisms used to moderate online speech suffer from a number of inherent deficiencies which render them a less-than-perfect tool for such a significant task. On the other hand, however, a lack of transparency and consistency on part of the platforms further exacerbates these problems.

As far as copyright claims are concerned, the DMCA does not require SMPs are to verify the substantive accuracy of takedown requests, which incentivises platforms to take claims by alleged rights-holders at face value and leaves users vulnerable to abuse.

In the case of violations of platform rules and policies, SMPs will review the merits of a reported violation, however, NTD compels content moderators to make a judgement without having heard both each sides of the story. Furthermore, since this decision-making process takes place in a black box, the outcome might not only be difficult to comprehend but may seem unjust and unfair even if it is in fact not.

The main problem is that Facebook, Twitter and Youtube are highly selective about what and how much they share of their review processes and become more secretive the further one seeks peer behind the scenes. This inconsistent approach has resulted in users having very little knowledge of how the SMPs decide over user-generated content. Furthermore, due to the lack of effective appeal mechanisms on Facebook, Twitter and Youtube, users may be forced to tolerate a penalty or judgement which they did not deserve in the first place.

The need to improve content moderation and the protection of users' rights as well as SMPs' role as public platforms and de facto courts for online speech makes it both necessary and appropriate for SMPs to adopt due process requirements.

By requiring SMPs to clearly and openly communicate their policy and enforcement decisions; apply their governing rules consistently and without bias; and allow users to present their

case and make appeals, due process will improve the protection of users' rights without unnecessary limitation of their freedoms, which is in the interest of users, SMPs and society as a whole.

The examples of hate speech and copyright illustrate additional problems in the enforcement of platform rules which must be taken into consideration when proposing reforms.
Due to the practical difficulty of bringing a copyright case to court, Youtube often ends up being the first and only arbiter that users involved in copyright disputes have access to. In the absence of an alternative forum to achieve justice, it is crucial that Youtube observes due process standards to ensure that users' rights are enforced and protected as best as possible.
The example of hate speech illustrated how Facebook's failure to regulate content in timely and consistent manner directly or indirectly facilitated the perpetration of violence. The dramatic real-life consequences of hate speech against the Rohingya people highlights the importance of due process not only for the content poster but for all other users and the platform itself.

5.2) Concluding remarks

In the aftermath of the Cambridge Analytica Scandal and the genocide in Myanmar, internet users and governments appear to be in agreement that the current content moderation system of Facebook, Twitter and Youtube is unsustainable and in urgent need of reform.
Even SMPs themselves have begun to look for new and more appropriate options for content moderation, such as Facebook's idea to set up an independent content moderation Supreme Court or oversight board.[106]

While such novel approaches should not be disregarded, they must not distract from the need to improve the current NTD-based system which is unlikely to disappear anytime soon since alternative models such as notice-and-staydown mechanisms present equally difficult procedural challenges. Therefore, SMPs should focus their efforts on addressing the

[106] Facebook, *Global Feedback & Input on the Facebook Oversight Board for Content Decisions* (2019) <https://fbnewsroomus.files.wordpress.com/2019/06/oversight-board-consultation-report-2.pdf> accessed 01/09/2019.

deficiencies of the current system and improving it as best as they can. As has been demonstrated throughout this paper, there is a need for due process both in theory and practice and it is an appropriate mechanism due to SMPs public function and role as judges of free speech.

However, before proposing concrete ways in which SMPs should make their content moderation processes more due process conforming, it is important to acknowledge some of the legal and practical restraints within which SMPs have to operate, as well as the need for regulatory reform at the international and national level.

As discussed in Chapter 4, using the DMCA as the statutory basis for copyright claims across the world is problematic and somewhat at odds with the non-US centric position taken by SMPs in respect of other prohibited behaviour.[107] At the same time, however, other legislative instruments such as the ECD contain no procedural guidelines which make them very difficult to implement. Furthermore, the high legal fees in the US which deter many users from pursuing litigation are beyond Facebook's or any other SMP's control.

As far as the timely assessment of, and correct response to, political events is concerned, this is a challenge that not only global companies like Facebook have faced. On the contrary, it is a problem that the international community at large has grappled with and often failed at for many years.[108]

Nonetheless, rather than serve as a potential excuse for SMPs' poor content moderation decisions, these factors demonstrate once again the need for SMPs to be transparent about the content moderation decisions they make and the challenges they face; to apply their rules consistently even at times of change and unrest; and to treat users equally and protect their right to present their case when they may not have another opportunity to do so.

[107] For instance, SMPs' rules against hate speech are more European in their approach in that they do not require content to incite "imminent lawless action" to justify its removal, as would be the case under US law: *Brandenburg v Ohio*, 395 U.S. 444 (1969).

[108] e.g. the UN's failure to intervene in the 1994 Rwandan Genocide or the failed UN peacekeeping mission in Srebrenica.

5.3) Suggestions

Based on the evaluation of the different stages in Facebook, Twitter and Youtube's content moderation processes (*Chapter 3*), the following measures are recommended to make the three platforms' regulatory procedures more transparent, even and fair.

3.1) Platform rules and policies and transparency reports

Although the publication of Facebook, Twitter and Youtube's platform rules is an important step in the right direction, policy changes on the basis of real-life events should be made public to allow for scrutiny and feedback from experts and stakeholders. As the example of Myanmar has demonstrated, local stakeholders are often in better position to recommend the right course of action as well as predict the response to new policies.

Furthermore, to allow for a comprehensive comparison and analysis of the enforcement of platform rules and policies, it is vital that SMPs include all types of violations in their transparency reports and present the numbers and statistics in a standardised format to allow for a comparison between the three platforms (as envisaged by the Santa Clara principles).

3.2) Detecting and reporting violations

While the use of manual, NTD-based user reports is very well documented by all three platforms, there is a need for more transparency surrounding the use of automated systems, not least because SMPs increasingly rely on such technologies to detect infringing content.

In the area of copyright enforcement where mechanisms such as Content ID make up a considerable part of copyright claims, a better overview of the prevalence of alleged copyright infringements is particularly important. In the case of wrongful takedowns, users will want to know if the mistake was made by a machine or a human, since the solution/remedy will differ. Importantly, SMPs should provide detailed information as to how and when automated systems are used (pre- or post-upload, proactively or reactively), how accurate they are and how they are being improved.

3.3) Reviewing reported violations

To make the internal review process more transparent, platforms should inform users of when, how, by whom and on what basis reviews of reported violations are carried out.

While it may not be practical to always publish the constantly changing internal guidelines, platforms should openly acknowledge that content moderation decisions are made on the basis of rules which are not available to the public (at least, not in such detail).
This type of open communication about internal processes will help fight the impression that the content review system is arbitrary and could foster more understanding regarding the difficulty of content moderation.

3.4) Enforcement options

Any penalties that SMPs have at their disposal ought to be proportionate, nuanced to reflect the severity of the breach in question. To ensure that sanctions have a deterring effect, users must be made aware of what they will face if they breach a platform's rules.

With regards to enforcement, SMPs must ensure that they treat users equally and if any exceptions have to be made, they need to be transparent, justifiable and openly communicated. While Twitter's and Facebook's newsworthiness exception may be justified in the case of Donald Trump, who uses Twitter as if it were an official channel of political communication, it should not become the norm for all people with a public profile.

3.5) Appeals

Appeals of content moderation decisions must be available for all types of violations and sanctions and should be dealt with within a certain timeframe. Furthermore, information on appeals (how to make them, how they will be dealt with etc.) must be easier to access to for users, and not only be communicated once a violation has occurred. Importantly, platforms should offer users the opportunity to include additional information (e.g. defences) in their appeals (as offered by Twitter) to be able to present their case.
As with enforcement, SMPs need to pay closer attention to the equal treatment of users. Famous Youtubers should not receive preferential treatment when it comes to appealing copyright strikes simply because they have a large following.

6.) BIBLIOGRAPHY

Table of cases

Canada

Baker v Canada (Minister of Citizenship and immigration) [1999] 2 R.C.S., Supreme Court of Canada, No. 25823.

United States

Brandenburg v Ohio, 395 U.S. 444 (1969)
Hosseinzadeh v Klein, 276 F. Supp. 3d 34 (S.D.N.Y. 2017) (h3h3 case)
Lenz v Universal Music Corp, 801 F.3d 1126 (2015), (9th Cir. 2015)

Table of legislation

European Union

European Parliament and Council, Directive 2000/31/EC on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce'), 8 June 2000.

European Parliament and Council, *Legislative resolution on the proposal for a directive of the European Parliament and of the Council on copyright in the Digital Single Market*, 26 March 2019 <http://www.europarl.europa.eu/doceo/document/TA-8-2019-0231_EN.html> accessed 09/06/2019.

United States

Digital Millennium Copyright Act (DMCA) 17 USC
Copyright Act of 1976 17 USC

Other sources

A

Adena M, Enikolopov R, Petrova M, Santarosa V and Zhuravskaya E, 'Radio and the Rise of The Nazis in Prewar Germany' (2015) 130(4) The Quarterly Journal of Economics 1885.

Alexander J, 'Youtubers and record lables are fighting, and record labels keep winning' (*The Verge*, 24 May 2019) <https://www.theverge.com/2019/5/24/18635904/copyright-youtube-creators-dmca-takedown-fair-use-music-cover> accessed 19/08/2019.

Angelopoulos C and Smet S, 'Notice-and-fair-balance: how to reach a compromise between fundamental rights in European intermediary liability' (2016) 8(2) Journal of Media Law 266.

Asher Hamilton I, 'Twitter wants to label tweets from public figures that break its rules — and even Trump could be named and shamed' (*Business Insider*, 29 March 2019) <https://www.businessinsider.com/twitter-to-label-tweets-from-public-figures-like-trump-that-violate-rules-2019-3?r=US&IR=T> accessed 23/08/2019.

B

Bailey J, 'YouTube's Copyright Problem' (*Plagiarism Today*, 23 October 2013) <https://www.plagiarismtoday.com/2013/10/23/youtubes-copyright-problem/> accessed 19/08/2019.

Balkin J M, 'Free Speech is a Triangle' (2018) 118(7) Columbia Law Review 2011.

Bartholomew T B, 'The Death of Fair Use in Cyberspace: Youtube and the Problem with Content ID' 13(1) Duke Law & Technology Review 66.

The Berkman Klein Center for Internet & Society at Harvard University, *Lumen Database* (2017) <https://lumendatabase.org> accessed 01/09/2019.

BSR, *Human Rights Impact Assessment: Facebook in Myanmar* (October 2018) <https://fbnewsroomus.files.wordpress.com/2018/11/bsr-facebook-myanmar-hria_final.pdf> accessed 24/08/2019.

C

Carlisle S, 'DMCA "Takedown" Notices: Why "Takedown" Should Become "Take Down and Stay Down" and Why It's Good for Everyone' (*Nova Southeastern University*, 23 July 2014) <http://copyright.nova.edu/dmca-takedown-notices/> accessed 03/05/2019.

The Cato Institute, 'Free Speech in an Age of Social Media' (*Youtube*, 19 May 2019) <https://youtu.be/agdZCzbweqo> accessed 20/06/2019.

Chen A, 'The laborers who keep dick pics and beheadings out of your Facebook feed' (*Wired*, 23 October 2014) <https://www.wired.com/2014/10/content-moderation/> accessed 17/07/2019.

Cohen N, 'YouTube Is Purging Copyrighted Clips' *The New York Times* (New York, 30 October 2006) <https://www.nytimes.com/2006/10/30/technology/30youtube.html> accessed 19/08/2019.

Crawford K and Gillespie T, 'What is a flag for? Social media reporting tools and the vocabulary of complaint' (2014) 18(3) New Media & Society 410.

D

Dwoskin E, 'YouTube's arbitrary standards: Stars keep making money even after breaking the rules' *The Washington Post* (Washington DC, 9 August 2019) <https://www.washingtonpost.com/technology/2019/08/09/youtubes-arbitrary-standards-stars-keep-making-money-even-after-breaking-rules/> accessed 01/09/2019.

E

EFF, *Takedown Hall of Shame* (2019) <https://www.eff.org/takedowns/> accessed 01/09/2019.

European Digital Rights, *#SaveYourInternet* (2019) <https://saveyourinternet.eu> accessed 10/06/2019.

F

Facebook, *Appealing a Claim of Copyright Infringement Made Under the DMCA (Counter-Notification)* (2019) <https://www.facebook.com/legal/copyright.php?howto_appeal=1> accessed 01/09/2019.

Facebook Community forum, 'Heather's question' (10 October 2018) <https://m.facebook.com/help/community/question/?id=158432015100013&rdrhc> accessed 18/07/2019.

Facebook Community forum, 'Julie's question' (16 August 2018) < https://m.facebook.com/help/community/question/?id=10215596085529558&rdrhc> accessed 18/07/2019.

Facebook Community forum, 'Francie's question' (17 August 2018) < https://m.facebook.com/help/community/question/?id=1726991954016707&answer_id=1744943225554913> accessed 18/07/2019.

Facebook, *Community Standards* (2019) <https://www.facebook.com/communitystandards/> accessed 09/07/2019.

Facebook, 'Community Standards Enforcement Report' *Facebook Transparency Report* (2019) <https://transparency.facebook.com/community-standards-enforcement> accessed 01/07/2019.

Facebook, 'Enforcing Our Community Standards' (*Facebook Newsroom,* 6 August 2018) <https://newsroom.fb.com/news/2018/08/enforcing-our-community-standards/> accessed 17/07/2019.

Facebook, *Facebook Terms and Policies* (2019) <https://en-gb.facebook.com/policies> accessed 01/09/2019.

Facebook, *Facebook Transparency Report* (2019) <https://transparency.facebook.com> accessed 01/09/2019.

Facebook, *Global Feedback & Input on the Facebook Oversight Board for Content Decisions* (2019) <https://fbnewsroomus.files.wordpress.com/2019/06/oversight-board-consultation-report-2.pdf> accessed 01/09/2019.

Facebook, 'Intellectual Property' *Facebook Transparency Report* (2019) <https://transparency.facebook.com/intellectual-property> accessed 24/08/2019.

Facebook, *Rights Manager* (2019) <https://rightsmanager.fb.com> accessed 01/09/2019.

Facebook, *Understanding the Community Standards Enforcement Report* (2019) <https://transparency.facebook.com/community-standards-enforcement/guide#section4> accessed 01/09/2019.

G

Google, 'Human flags by flagging reasons' *Youtube Community Guidelines enforcement report* (2019) <https://transparencyreport.google.com/youtube-policy/flags> accessed 23/08/2019

Google, *YouTube Community Guidelines enforcement* (2019) <https://transparencyreport.google.com/youtube-policy/removals?hl=en> accessed 09/07/2019.

H

h3h3Productions, 'We're Still Being Sued' (27 February 2017) <https://www.youtube.com/watch?v=m40bWgWH8Ro> accessed 20/08/2019.

Harmon E, 'Don't Sacrifice Fair Use to the Bots' (*EFF*, 1 March 2019) <https://www.eff.org/deeplinks/2019/03/dont-sacrifice-fair-use-bots> accessed 19/08/2019.

Hollander-Blumoff R and Tyler T R, 'Procedural Justice and the Rule of Law: Fostering Legitimacy in Alternative Dispute Resolution' (2011) Journal of Dispute Resolution.

Hootsuite and We Are Social, *Digital 2019 Global Digital Overview* (2019) <https://datareportal.com/reports/digital-2019-global-digital-overview> accessed 31/08/2019.

J

Johnson Swan K, 'United States: The True Cost Of Defending Against Copyright Infringement Litigation' (*Mondaq*, 19 August 2015) <http://www.mondaq.com/unitedstates/x/421188/Copyright/The+True+Cost+Of+Defending+Against+Copyright+Infringement+Litigation> accessed 01/09/2019.

K

Kadri T and Klonick K, 'Facebook v. Sullivan: Public Figures and Newsworthiness in Online Speech' (2019) Southern California Law Review (Forthcoming), St. John's Legal Studies Research Paper No. 19-0020 <http://dx.doi.org/10.2139/ssrn.3332530> accessed 21/08/2019.

Keats Citron D and Norton H, 'Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age' (2011) 91 Boston University Law Review L REV 1435.
Klonick K, 'The new governors: the people, rules, and processes governing online speech' (2018) 131 Harvard Law Review 1598.

Koebler J and Cox J, 'The Impossible Job: Inside Facebook's Struggle to Moderate Two Billion People' (*Vice*, 23 August 2018) <https://www.vice.com/en_us/article/xwk9zd/how-facebook-content-moderation-works> accessed 01/09/2019.

Kosoff M, 'Facebook's Hate-speech problem may be bigger than it realized' (*Vanity Fair,* 21 August 2018) <https://www.vanityfair.com/news/2018/08/facebooks-hate-speech-problem-may-be-bigger-than-it-realized> accessed 24/08/2019.

Kyl J, *Covington Interim Report* (2019) <https://fbnewsroomus.files.wordpress.com/2019/08/covington-interim-report-1.pdf> accessed 24/08/2019.

L

Laub Z, 'Hate Speech on Social Media: Global Comparisons' (*Council on Foreign Affairs*, 7 June 2019) <https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons> accessed 24/08/2019.

M

Manila Principles on Intermediary Liability (2015) <https://www.eff.org/files/2015/10/31/manila_principles_1.0.pdf> accessed 23/08/2019.

Marshall P D, 'A comparative analysis of the right to appeal' (2011) 22(1) Duke Journal of Comparative & International Law 1.

Maxeiner J R, 'Some Realism about Legal Certainty in the Globalization of the Rule of Law' (2008) 31(1) Houston Journal of International Law 27.

McAuley J, 'France moves toward a law requiring Facebook to delete hate speech within 24 hours' *The Washington Post* (Washington DC, 9 July 2019) <https://www.washingtonpost.com/world/europe/france-moves-toward-a-law-requring-facebook-to-delete-hate-speech-within-24-hours/2019/07/09/d43b24c2-a25d-11e9-a767-d7ab84aef3e9_story.html> accessed 24/08/2019.

McSherry C, 'After More Than a Decade of Litigation, the Dancing Baby Has Done His Part to Strengthen Fair Use for Everyone' (*EFF*, 27 June 2018) <https://www.eff.org/deeplinks/2018/06/after-more-decade-litigation-dancing-baby-ready-move> accessed 19/08/2019.

Mostert F, 'Free Speech and Internet Regulation' (2019) Journal of Intellectual Property Law & Practice.

Mostert F and Lambert J, 'Study on IP enforcement measures, especially anti-piracy measures in the digital environment' (WIPO Advisory Committee on Enforcement, 14th Session 2-4 September, Draft paper 10 May 2019).

Müller K and Schwarz C, *Fanning the Flames of Hate: Social Media and Hate Crime* (University of Warwick CAGE Working Paper Series No.373, May 2018) <https://warwick.ac.uk/fac/soc/economics/research/centres/cage/manage/publications/373-2018_schwarz.pdf> accessed 21/08/2019.

Mueller M, *Challenging the Social Media Moral Panic* (Cato Institute Policy Analysis No. 876, 23 July 2019) <https://object.cato.org/sites/cato.org/files/pubs/pdf/pa-876-update.pdf> accessed 21/08/2019.

Myanmar Civil Society Organizations, *Open Letter to Mark Zuckerberg* (5 April 2018) <https://drive.google.com/file/d/1Rs02G96Y9w5dpX0Vf1LjWp6B9mp32VY-/view> accessed 24/08/2019.

N

Newton C, 'The Trauma Floor - The secret lives of Facebook moderators in America' (*The Verge*, 25 February 2019) <https://www.theverge.com/2019/2/25/18229714/cognizant-

facebook-content-moderator-interviews-trauma-working-conditions-arizona> accessed 16/06/2019.

P

Pariser E, *The Filter Bubble: What the Internet is hiding from you* (Penguin Books 2011).

Persak N, 'Procedural Justice Elements of Judicial Legitimacy and their Contemporary Challenges' (2016) 6(3) Onati Socio-legal Series 749.

Plovanic J, 'YouTube (Still) Has a Copyright Problem' Washington Journal of Law, Technology & Arts (*WJLTA Blog*, 28 February 2019) <https://wjlta.com/2019/02/28/youtube-still-has-a-copyright-problem/> accessed 19/08/2019.

R

Rosen J, 'The Delete Squad: Google, Twitter, Facebook and the new global battle over the future of free speech' (*The New Republic*, 29 April 2013) <https://newrepublic.com/article/113045/free-speech-internet-silicon-valley-making-rules> accessed 31/08/2019.

S

Safi M, 'Myanmar treatment of Rohingya looks like 'textbook ethnic cleansing', says UN' *The Guardian* (London, 11 September 2017) <https://www.theguardian.com/world/2017/sep/11/un-myanmars-treatment-of-rohingya-textbook-example-of-ethnic-cleansing> accessed 22/08/2019.

Samples J, *Why the Government Should Not Regulate Content Moderation of Social Media* (Cato Institute Policy Analysis No. 865, 9 April 2019) <https://object.cato.org/sites/cato.org/files/pubs/pdf/pa_865.pdf> accessed 21/08/2019.

Samples J, *False Assumptions Behind the Current Drive to Regulate Social Media* (Cato Institute Blog, 23 July 2019) <https://www.cato.org/blog/false-assumptions-behind-current-drive-regulate-social-media> accessed 21/08/2019.

Sands M, 'Why Copyright Will Be The Biggest Issue For Youtube In 2019 (Updated)' (*Forbes*, 30 December 2018) <https://www.forbes.com/sites/masonsands/2018/12/30/why-copyright-will-be-the-biggest-issue-for-youtube-in-2019/#200c176b1c12> accessed 19/08/2019.

The Santa Clara Principles On Transparency and Accountability in Content Moderation (2018) <https://santaclaraprinciples.org> accessed 24/08/2019.

Stecklow S, *Why Facebook is losing the war on hate speech in Myanmar* (Reuters Special report, 15 August 2018) <https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/> accessed 24/08/2019.

T

TeamYouTube[Help], 'The Life of a Flag' (*Youtube*, 23 April 2018) <https://www.youtube.com/watch?v=WK8qRNSmhEU> accessed 11/07/2019.

Thomas E, 'Facebook Keeps Failing in Myanmar' (*Foreign Policy*, 21 June 2019) <https://foreignpolicy.com/2019/06/21/facebook-keeps-failing-in-myanmar-zuckerberg-arakan-army-rakhine/> accessed 20/08/2019.

Townsend B, 'James Charles and his fans win copyright battle against YouTube' (*We The Unicorns*, 9 April 2019) <https://www.wetheunicorns.com/youtubers/james-charles/james-charles-fans-win-copyright-battle/> accessed 19/08/2019.

Trendacosta K, 'YouTube's New Lawsuit Shows Just How Far Copyright Trolls Have to Go Before They're Stopped' (*EFF*, 21 August 2019) <https://www.eff.org/deeplinks/2019/08/youtubes-new-lawsuit-shows-just-how-far-copyright-trolls-have-go-theyre-stopped> accessed 01/09/2019.

Twitter, 'Copyright notices' *Transparency report* (2019) <https://transparency.twitter.com/en/copyright-notices.html> accessed 23/08/2019.

Twitter, 'Copyright policy' *Twitter Rules and policies* (2019) <https://help.twitter.com/en/rules-and-policies/copyright-policy> accessed 01/09/2019.

Twitter, *Our approach to policy development and enforcement philosophy* (2019) <https://help.twitter.com/en/rules-and-policies/enforcement-philosophy> accessed 17/07/2019.

Twitter, *Sensitive media policy* (2019) <https://help.twitter.com/en/rules-and-policies/media-policy> accessed 01/07/2019.

Twitter, *Transparency Report* (2019) <https://transparency.twitter.com> accessed 01/09/2019.

Twitter, *The Twitter Rules* (2019) <https://help.twitter.com/en/rules-and-policies/twitter-rules> accessed 29/08/2019.

Twitter, *Twitter Rules and policies* (2019) <https://help.twitter.com/en/rules-and-policies#general-policies> accessed 01/09/2019.

Twitter, 'Twitter Rules enforcement' *Transparency Report* (2019) <https://transparency.twitter.com/en/twitter-rules-enforcement.html> accessed 24/08/2019.

Twitter Safety, 'Introduction of in-app appeal mechanism' (*Twitter*, 2 April 2019) <https://twitter.com/TwitterSafety/status/1113139073303089152> accessed 11/07/2019.

Tyler T R, 'Procedural Justice, Legitimacy, and the Effective Rule of Law' (2003) 30 Crime and Justice 283.

Tyler T R, *Why People obey the Law* (Princeton University Press 2007).

U

United Nations Human Rights Council, *Report of the independent international fact-finding mission on Myanmar* (A/HRC/39/64, 12 September 2018) <https://www.ohchr.org/Documents/HRBodies/HRCouncil/FFM-Myanmar/A_HRC_39_64.pdf> accessed 24/08/2019.

The University of Houston Law Center, 'Racists, Bigots and the Law on the Internet', seminar hosted by the University of Houston Law Center and Anti-Defamation League (ADL), 4 October 2012 (*Youtube*, 10 October 2012) <https://youtu.be/aqqvYPyr6cI> accessed 10/07/2019.

Urban J M, Karaganis J and Schofield B L, *Notice and Takedown in Everyday Practice* (University of California, Berkeley and Columbia University 2017).

W

Walzel S, 'European Commission Consults on Notice and Takedown' (*LSE Media Policy Project Blog*, 24 August 2012) <(http://eprints.lse.ac.uk/78705/1/European%20Commission%20Consults%20on%20Notice%20and%20Takedown%20_%20LSE%20Media%20Policy%20Project.pdf> accessed 03/06/2019.

Warofka A (Facebook Product Policy Manager), 'An Independent Assessment of the Human Rights Impact of Facebook in Myanmar' (*Facebook Newsroom*, 5 November 2018) <https://newsroom.fb.com/news/2018/11/myanmar-hria/> accessed 10/07/2019.

Wolfson S, ‘Facebook labels declaration of independence as 'hate speech’’ *The Guardian* (London, 5 July 2018) <https://www.theguardian.com/world/2018/jul/05/facebook-declaration-of-independence-hate-speech> accessed 10/07/2019.

Wong J C, ‘'Overreacting to failure': Facebook's new Myanmar strategy baffles local activists’ *The Guardian* (London, 7 February 2019) <https://www.theguardian.com/technology/2019/feb/07/facebook-myanmar-genocide-violence-hate-speech> accessed 24/08/2019.

Y

Yanagizawa-Drott D, ‘Propaganda and conflict: evidence from the Rwandan Genocide’ (2014) 129(4) Quarterly Journal of Economics 1947.

York J C and McSherry C, ‘Content Moderation is Broken. Let Us Count the Ways.’ (*EFF*, 29 April 2019) <https://www.eff.org/deeplinks/2019/04/content-moderation-broken-let-us-count-ways> accessed 01/09/2019.

Youtube, *Community Guidelines* (2019) <https://www.youtube.com/intl/en-GB/yt/about/policies/#community-guidelines> accessed 29/08/2019.

Youtube, *Community Guideline strikes basics* (2019) <https://support.google.com/youtube/answer/2802032?hl=en> accessed 23/08/2019.

Youtube, *Copyright counter notification basics* (2019) <https://support.google.com/youtube/answer/2807684?hl=en-GB&ref_topic=9282678> accessed 01/09/2019.

Youtube, *Copyright Infringement Notification* (2019) <https://www.youtube.com/copyright_complaint_form> accessed 24/08/2019.

Youtube, *How Content ID works* (2019) <https://support.google.com/youtube/answer/2797370> accessed 24/08/2019.

Youtube, *Strikes FAQ* (2019) <https://support.google.com/youtube/answer/9235777?hl=en&ref_topic=2803138> accessed 23/08/2019

Youtube, *The Importance of Context* (2019) <https://support.google.com/youtube/answer/6345162?hl=en> accessed 17/07/2019.

Youtube, ‘Updates to our manual Content ID claiming policies’ (Y*outube Creator Blog*, 15 August 2019) <https://youtube-creators.googleblog.com/2019/08/updates-to-manual-claiming-policies.html> accessed 01/09/2019.

Youtube, *Youtube v Christopher L Brady: Demand for Jury Trial* (United States District Court District of Nebraska, Case No. 19-353, 19 August 2019) <https://torrentfreak.com/images/Youtube-v-Christopher-Brady-DMCA-abuse-complaint-191908.pdf> accessed 01/09/2019.

Youtube, *Youtube Policies* (2019) <https://support.google.com/youtube/topic/2803176?hl=en-GB&ref_topic=6151248,3230811,3256124,> accessed 01/09/2019.

Z

Zuckerberg M, *Witness Testimony* (Facebook, Social Media Privacy, and the Use and Abuse of Data: Hearing before the United States Senate Committee on the Judiciary and the United States Senate Committee on Commerce, Science and Transportation, 10 April 2018) <https://www.judiciary.senate.gov/imo/media/doc/04-10-18%20Zuckerberg%20Testimony.pdf> accessed 24/08/2019.

7.) APPENDIX: The Santa Clara Principles

THE SANTA CLARA PRINCIPLES

On Transparency and Accountability in Content Moderation[109]

These principles are meant to serve as a starting point, outlining minimum levels of transparency and accountability that we hope can serve as the basis for a more in-depth dialogue in the future.

On the occasion of the first Content Moderation at Scale conference in Santa Clara, CA on February 2nd, 2018, a small private workshop of organizations, advocates, and academic experts who support the right to free expression online was convened to consider how best to obtain meaningful transparency and accountability around internet platforms' increasingly aggressive moderation of user-generated content.
Now, on the occasion of the second Content Moderation at Scale conference in Washington, DC on May 7th, 2018, we propose **these three principles** as initial steps that companies engaged in content moderation should take to provide meaningful due process to impacted speakers and better ensure that the enforcement of their content guidelines is fair, unbiased, proportional, and respectful of users' rights.

- ACLU Foundation of Northern California
- Center for Democracy & Technology
- Electronic Frontier Foundation
- New America's Open Technology Institute
- Irina Raicu, Markkula Center for Applied Ethics, Santa Clara University
- Nicolas Suzor, Queensland University of Technology
- Sarah T. Roberts, Department of Information Studies, School of Education & Information Studies, UCLA
- Sarah Myers West, USC Annenberg School for Communication and Journalism

[109] *The Santa Clara Principles on Transparency and Accountability in Content Moderation* (2018) <https://santaclaraprinciples.org> accessed 30/08/2019 (n37).

1. NUMBERS

Companies should publish the numbers of posts removed and accounts permanently or temporarily suspended due to violations of their content guidelines.

At a minimum, this information should be broken down along each of these dimensions:

- Total number of discrete posts and accounts flagged.
- Total number of discrete posts removed and accounts suspended.
- Number of discrete posts and accounts flagged, and number of discrete posts removed and accounts suspended, by category of rule violated.
- Number of discrete posts and accounts flagged, and number of discrete posts removed and accounts suspended, by format of content at issue (e.g., text, audio, image, video, live stream).
- Number of discrete posts and accounts flagged, and number of discrete posts removed and accounts suspended, by source of flag (e.g., governments, trusted flaggers, users, different types of automated detection).
- Number of discrete posts and accounts flagged, and number of discrete posts removed and accounts suspended, by locations of flaggers and impacted users (where apparent).

This data should be provided in a regular report, ideally quarterly, in an openly licensed, machine-readable format.

2. NOTICE

Companies should provide notice to each user whose content is taken down or account is suspended about the reason for the removal or suspension.

In general, companies should provide detailed guidance to the community about what content is prohibited, including examples of permissible and impermissible content and the guidelines used by reviewers. Companies should also provide an explanation of how automated detection is used across each category of content. When providing a user with

notice about why her post has been removed or an account has been suspended, a minimum level of detail for an adequate notice includes:

- URL, content excerpt, and/or other information sufficient to allow identification of the content removed.
- The specific clause of the guidelines that the content was found to violate.
- How the content was detected and removed (flagged by other users, governments, trusted flaggers, automated detection, or external legal or other complaint). The identity of individual flaggers should generally not be revealed, however, content flagged by government should be identified as such, unless prohibited by law.
- Explanation of the process through which the user can appeal the decision.

Notices should be available in a durable form that is accessible even if a user's account is suspended or terminated. Users who flag content should also be presented with a log of content they have reported and the outcomes of moderation processes.

3. APPEAL

Companies should provide a meaningful opportunity for timely appeal of any content removal or account suspension.

Minimum standards for a meaningful appeal include:

- Human review by a person or panel of persons that was not involved in the initial decision.
- An opportunity to present additional information that will be considered in the review.
- Notification of the results of the review, and a statement of the reasoning sufficient to allow the user to understand the decision.

In the long term, independent external review processes may also be an important component for users to be able to seek redress.

Acknowledgements

We thank Santa Clara University's High Tech Law Institute for organizing the Content Moderation & Removal at Scale conference, as well as Eric Goldman for supporting the convening of the workshop that resulted in this document. That workshop was also made possible thanks to support from the Internet Policy Observatory at the University of Pennsylvania. Suzor is the recipient of an Australian Research Council DECRA Fellowship (project number DE160101542).