The Dickson Poon
School of Law

Legal Opinion

Practice Project Title: Legal Opinion on How the European Commission can Improve its Content Moderation Regulations to Best Mitigate Harms Created by ‘False Information’ on Social Media

Table of Contents

Executive Summary

To: The European Commission

1 September 2022

Social media platforms are the place to be seen and the place to be heard. It is also a place to like, share and comment on a variety of topics. It is where you see that your long-lost friend has finally gotten married to her high school sweetheart or your uncle's pictures of their new puppy, Billy. Unfortunately, it is now *the* most dominant place to find information about important events in society, whether the upcoming elections or the new Corona Virus (COVID-19) measures. Although probably not the platform's initial intention, that is now what social media is used for.

Hate speech, misinformation, and disinformation or 'fake news' contaminate social media platforms around the world and creates adverse effects on society, human rights, democracy, and user safety. Given how social media platforms operate, it is likely that those who believe in such false information will only see content that supports their own views in the digital space. Some platforms are so dominant that through their network effects and, as Eli Pariser called it, filter bubbles, users are subject to engaging with immensely dangerous information that could cause harm to themselves and others. Effective responses to protect society and users are needed at many levels ranging from formal laws to user empowerment. While it should not be up to one government or one social media platform to decide what is fact or fiction, we need to find a way to collaborate and put our minds together to protect users and create a safe online space that does not rely on profit maximisation or inappropriate use of data.

I, on behalf of the Electronic Frontier Foundation, address this report to the European Commission as it is your responsibility to ensure that any fundamental rights are enforced, and laws are initiated. Through the Digital Services Act (DSA), the European Commission is best equipped to ensure that content governance is redefined and enforced harmoniously across the European Union and that the digital space is made safer. The severity of harm and danger for users due to false information on social media platforms poses an imminent threat if not quickly and correctly approached. User protection is ever more at stake in the current landscape of social media platforms and content moderation.

This report provides background to the current pressing issues of content moderation, the role of social media and Artificial Intelligence, lack of coherence in definitions, the surge in the volume of content, and the current regulatory approaches to content moderation. The report aims to illustrate the severity of mismatched content moderation, the urgency of amendments for user protection and content moderation regulation beyond the DSA and the polluting effect false information will continue to play if no change is made. The report predominantly draws on examples from Facebook due to its prevalence and scrutiny, however, to only focus on Facebook is dangerously narrow given how hyper-connected platforms and users are. As such, several proposals will be raised and analysed with a call for external oversight.

I trust that together we can ensure that social media platforms are held liable for their role in the dissemination of false information and that transparency requirements are adequately enforced and that we create a genuinely safe and dependable digital space that protects all its users.

1. Content Moderation Generally

Content moderation has gained traction and has been heavily scrutinised in the last decade by consumers and civil society. The 2016 U.S. Election[1] and the COVID-19 pandemic have significantly compounded the need for an effective response to misinformation, disinformation, hate speech and illegal content.[2] With increased access to news through social media, a dystopic future where governments need to intervene entirely independently of platforms and users does not seem farfetched, especially as content moderation by platforms has been in discussion for far too long with little meaningful change. It has been too long because Facebook started back in 2004, Twitter in 2006, and new platforms like TikTok do not seem to have taken on board much of the discussion. Although, those who have made, arguably minimal efforts to improve their self-regulation would object to me asserting its ineffectiveness in alleviating the spread of harmful information. The need for better content moderation was yesterday.

Misinformation, fake news, disinformation, and conspiracy theories are all harmful pieces of media that are essentially "designed to mislead readers by looking and coming across as traditional media" and trustworthy".[3] Although each differs in its intent, for the purpose of this report, I will group them under the umbrella term of 'false information' as they all pose a severe danger to users' safety, democracy, incite violence, and adherence to measures and trust in others.

With today's use of social media platforms, it is easy for blatantly harmful content to become viral. However, the posts might not be viral because of the false misinformation. For example, Trump's posts could go viral because of the "popularity of Trump's account or the fact that he writes about politically charged subjects".[4] It is difficult to isolate the effect false information

[1] Hunt Allcott and Matthew Gentzkow, 'Social Media and Fake News in the 2016 Election' (2017) 31(1) Journal of Economic Perspectives 211, <https://web.stanford.edu/~gentzkow/research/fakenews.pdf> accessed 3 April 2022.

[2] Kasper Welbers and Michaël Opgenhaffen, 'Social Media Gatekeeping: An Analysis of The Gatekeeping Influence of Newspapers' Public Facebook Pages' (2018) 20(12) New Media & Society 4728, 4370 <https://doi.org/10.1177/1461444818784302> accessed 13 January 2022.

[3] Jay J. van Bavel and others, 'Using Social and Behavioural Science to Support COVID-19 Pandemic Response' (2020) 4(1) Nature Human Behaviour 460 <https://doi.org/10.1038/s41562-020-0884-z> accessed 9 April 2022.

[4] Chris Meserole, 'How Misinformation Spreads on Social Media – and What to do About It' (*Brookings*, 2018) <https://www.brookings.edu/blog/order-from-chaos/2018/05/09/how-misinformation-spreads-on-social-media-and-what-to-do-about-it/> accessed 14 August 2022.

has on users, but it is critical that users are protected from such hate speech, violence, and abuse.

To combat the spread of such information, platforms rely on content moderation which is the practice of checking user-generated content, whether an image, video, or text. Content moderation is central to social media platforms for several reasons. It is the primary mechanism for platforms to ensure user compliance with their community standards which aim to balance a place of expression compliant with the law– ultimately acting as a mechanism to protect users from harm and abuse. However, most importantly, it is a way for, say, Facebook to retain their users on their platform and engaging with content.[5]

A users' uploaded content may trigger moderation flags by site moderators or external parties for inappropriate content, leading to reviews by professional or technological moderators trained via Artificial Intelligence (AI).[6] For instance, an insensitive post or comment on Facebook will be flagged for review, and a moderator will request the user to remove the content. If the user continues to violate guidelines, the account may be blocked for a certain amount of time.[7] The same measures are in place if a post or comment violates the community standards.[8]

Content moderation is not as straightforward in practice, and a plethora of false information remains on the platform. To put it into perspective, roughly 240,000 photos alone are uploaded to Facebook every minute.[9] To sift through all the content is time-consuming and complicated before even sorting the bad information between more nuanced pieces or mere opinions. Even more so, content that catches virality will be more likely to be flagged by software, algorithms, and users and, thus, removed.[10] Some are harder to find. The endless possibility of harmful and

[5] Sarah T Roberts, 'Understanding Content Moderation', *Behind the Screen: Content Moderation in the Shadows of Social Media* (1st edn, Yale University Press 2019) <https://www.jstor.org/stable/j.ctvhrcz0v.5> accessed 30 July 2022.

[6] Ibid.

[7] Kyle Langvardt, 'Regulating Online Content Moderation' (2017) 106(5) Georgetown Law Journal 1353, 1355 <https://www.law.georgetown.edu/georgetown-law-journal/wp-content/uploads/sites/26/2018/07/Regulating-Online-Content-Moderation.pdf> accessed 4 April 2022.

[8] Facebook, 'Hate Speech' (*Facebook*, 2022) <https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/> accessed 12 March 2022; Langvardt (n 7).

[9] DOMO, 'Data Never Sleeps 9.0' (*DOMO*, 2021) <https://web-assets.domo.com/blog/wp-content/uploads/2021/09/data-never-sleeps-9.0-1200px-1.png> accessed 15 July 2022.

[10] Robert Gorwa, Reuben Binns and Christian Katzenbach, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance' (2020) 7 (1) Big Data and Society 1, 10 <https://doi.org/10.1177/2053951719897945> accessed 16 May 2022.

false information being severely amplified by content recommendation systems for higher visibility of content raises concerns about user safety.

One of the additional challenges of content moderation is that the nature of the content and its meanings differ within specific contexts. Even though AI, software and moderators can pick up on clear-cut phrases and words, it cannot best determine what is false or not within different contexts, cultures, and norms outside of its training environment. Due to its inherent inability to identify all the nuances, platforms and their AI should not become the arbiters of truth.[11] An ongoing pressing question already in the discussion is precisely who should be the ultimate arbiter of truth and who is in the best position to decide what is right and wrong?[12]

The leading dilemma facing content moderation is then whether false information should be entirely silenced and aggressively removed or to leave all information on the platform.[13] Both can severely impact the right to freedom of expression. As some say, one misstep and we may be facing censorship.[14] Therefore, companies and governments must primarily answer how social media platforms can respect users' freedom of expression and human rights while creating a safe online space.

In theory, filtering information is a good thing and can lead to less amplification of terrorist or other illegal content. Filtering also allows platforms to tailor content to each user and entice them to stay on the platform and use their products and services. However, it also means that users are bound to what the platform wants them to see and limits the digestion of information that might be relevant and safe outside of that 'bubble'.[15]

[11] Jason Pielemeier, 'Disentangling Disinformation: What Makes Regulating Disinformation So Difficult?' (2020) 4(1) Utah Law Review 917, 925 <https://dc.law.utah.edu/ulr/vol2020/iss4/1> accessed 14 March 2022.

[12] Pranav Rastogi, 'Should Social Media Platforms be the Arbitrator of Truth?' (*Medium*, 2021) <https://medium.com/redhill-review/should-social-media-platforms-be-the-arbitrator-of-truth-760df94baa70> accessed 13 August 2022.

[13] United Nations Human Rights Office of the High Commissioner, 'Moderating Online Content: Fighting Harm or Silencing Dissent?' (*OHCHR*, 2021) <https://www.ohchr.org/en/stories/2021/07/moderating-online-content-fighting-harm-or-silencing-dissent> accessed 19 July 2022.

[14] Pielemeier (n 11), 925.

[15] For example, suppose your Facebook friends, news sources, and other platforms all share the same opinion. An algorithm knows that and puts you into a filter bubble where you are less likely to be exposed to information that could challenge or broaden your worldview. Eli Pariser, 'Beware Online "Filter Bubbles"' <https://www.ted.com/talks/eli_pariser_beware_online_filter_bubbles> accessed 17 July 2022.

This multifaceted problem makes it especially difficult to find an appropriate way to ensure consumer protection while limiting the dominant power platforms can have on fundamental rights. Current content moderation practices, however, also raise concerns about freedom of expression, fundamental human rights, and democracy. Clarity and parameters are precisely what this report seeks to offer.

1.1. The Role of Social Media

It is without a doubt that social media has become *the* place to find information and plays a facilitative role in spreading information. The set-up of platforms like Facebook and Twitter allows information to be shared, consumed, and interacted with at vast rates. Today, these platforms play a vital role in the personal, political, and cultural life of billions of people.[16]

The ease and attractiveness of these platforms allows users to create their own profiles, connect with friends all around the globe, share updates or images and interact with other content they find interesting, and all with one click. The ease of access and intuitive use means that, for many, social media has frankly become an addiction[17] and a primary source for news and other information.[18] It would be naïve to think that only social media platforms act as a forum for false information. However, the issue of false information goes further and is present on unsuspecting platforms like Spotify.[19] The problem of false information and content moderation goes far beyond previous discussions that were centralised to social media and needs to be tackled quickly and appropriately.

Although false information is not a phenomenon of the digital era, it is much easier to come across false information, whether intentionally or not. The ease in access is due to the growth in the variety of content that can share false information, which goes beyond mere text and can

[16] José Van Dijck, Thomas Poell and Martijn De Waal, *The Platform Society. Public Values In A Connective World* (1st edn, Oxford University Press 2018).
[17] Tim Wu, *The Attention Merchants* (1st edn, Knopf Publishing Group 2016), 351.
[18] Krysten Crawford, 'Stanford Study Examines Fake News and the 2016 Presidential Election | Stanford News' (*Stanford News*, 2017) <https://news.stanford.edu/2017/01/18/stanford-study-examines-fake-news-2016-presidential-election/> accessed 12 January 2022.
[19] Adriana Sosa and others, 'An Open Letter to Spotify' (*An Open Letter to Spotify*, 2021) <https://spotifyopenletter.wordpress.com/2022/01/10/an-open-letter-to-spotify/> accessed 17 January 2022.

include videos, GIFs, and audio. Hence, leaving more opportunities for harmful content to spread and be digested.[20]

Facebook, a gatekeeper and advertising company, can predict what a user will be interested in based on their interactions. Facebook, and others, understand that viral and provocative content will keep the user on the platform.[21] If this remains the underlying goal of each social media platform, then platforms are exploiting users more than protecting them despite any efforts to show the contrary. As a result, social media platforms can influence users and society on a larger scale. It is exactly this mix of platform design, role as 'attention merchants' and surveillance that allows false information to thrive in the digital space.[22]

The mere fact that harmful information that most would instantly find mundane, such as ingesting bleach to cure yourself from COVID-19, can easily become viral and accepted, already causes concern.[23] The COVID-19 pandemic reminded civil society, regulators, and platforms how misinformation could spread as quickly as the virus itself if left unattended. The pandemic showed that one of the largest concerns regarding social media platforms is that their content moderation practices are inadequate against such information that can jeopardise the public health response when urgently needed. Without external influence, they remain potent gatekeepers of information.

Numerous factors highlight the problems with successful content moderation. Recent major events have emphasised that social media platforms, with immense global footprints, are not appropriately handling the challenges of content moderation and false information. Their shortcomings stress the need for more rigorous legal intervention should they want to continue

[20] Lisa Fazio, 'Out-Of-Context Photos Are A Powerful Low-Tech Form Of Misinformation' (*The Conversation,* 2020) <https://theconversation.com/out-of-context-photos-are-a-powerful-low-tech-form-of-misinformation-129959> accessed 13 June 2022; Manu Aggarwal, Abhijnan Dasgupta and Aakash Jaiswal, 'Safeguarding Social Media: How Effective Content Moderation Can Help Clean Up the Internet' (*Everest Group,* 2021) <https://www.everestgrp.com/safeguarding-social-media-how-effective-content-moderation-can-help-clean-up-the-internet-blog.html> accessed 25 June 2022.

[21] Hassan Salman, 'Regulating the Digital Resonance' (*American University Washington College of Law,* 2021) <https://digitalcommons.wcl.american.edu/cgi/viewcontent.cgi?article=1053&context=stu_upperlevel_papers> accessed 21 March 2022, 8.

[22] Wu (n 17).

[23] Matt Perez, 'Trump Suggests Injecting Coronavirus Patients With Light Or Disinfectants, Alarming Experts' (*Forbes,* 2020) <https://www.forbes.com/sites/mattperez/2020/04/23/trump-suggests-injecting-coronavirus-patients-with-light-or-disinfectants-contradicting-experts/?sh=37c041e04088> accessed 3 August 2022.

operating as an integral part of society. This report provides proposals for the various issues with content moderation and analyses social media’s failed attempts to alleviate the challenges.

2. Problems with Inconsistent Definitions

I have thus far referred to both false information and illegal content, yet the two are not entirely the same. Illegal content including content of a sexual nature or terroristic content are much clearer to identify than mis- and disinformation as its purpose is to target specific “individuals by instilling fear” or causing actual harm.[24] From the types of ‘false information’, misinformation is often harmful but legal and does not explicitly fit into a category as illegal content does. Mis- and disinformation are often not as easy to identify, and its harm is more disguised. Although misinformation does not always have to be illegal, it can be extremely harmful to society and can damage democratic elections,[25] decrease trust in public health measures[26] and can lead to uncalled violence.[27] Recent examples like the Capitol storming and the pandemic have shown how detrimental misinformation can be to public safety and how easy it is to believe and act on false information.

The European Commission has defined online disinformation as “verifiably false or misleading information that is created, presented, and disseminated for economic gain or to intentionally deceive the public, and may cause public harm”, and misinformation having no harmful intent.[28] Whether this encompasses false information that was only re-shared, liked or commented on by users is unclear despite showing up on people’s news feeds and having harmful effects.

While it would be tempting to continue to group them together within a regulatory regime, it would be wise to consider both mis- and disinformation individually due to their prevalence and imperceptible nature. However, a basic regulatory regime incorporating even this type of content will provide a stepping-stone for more specific regulation.

[24] Pielemeier (n 11), 923.
[25] Allcott (n 1).
[26] Van Dijck (n 16).
[27] BBC, 'Covid: Huge Protests Across Europe Over New Restrictions' (*BBC*, 2021) <https://www.bbc.co.uk/news/world-europe-59363256> accessed 11 March 2022.
[28] Alexandre de Streel et al, ‘Online Platforms' Moderation of Illegal Content Online Law, Practices and Options for Reform’ (*European Parliament Committee on Internal Market and Consumer Protection*, 2020) <https://www.europarl.europa.eu/RegData/etudes/STUD/2020/652718/IPOL_STU(2020)652718_EN.pdf> accessed 24 March 2022, 18; Martina Furlan, 'The New Code of Practice on Disinformation: An Attempt to Restore Responsibility in the Online Public Sphere' (89 Initiative, 2021) <https://89initiative.com/the-new-code-of-practice-on-disinformation-an-attempt-to-restore-responsibility-in-the-online-public-sphere/> accessed 23 March 2022.

More specifically, the definition of illegal content under European Union (EU) Law as per the DSA is scattered and relatively limited. Moreover, the definition is limited because illegal content is specified by European Union Law or national law.[29] The categories of illegal content include:

1. Child sexual abuse
2. Illegal hate speech like racist and xenophobic hate speech
3. Terrorist content
4. Commercial scams and frauds
5. Breaches of Intellectual Property Rights[30]

The fact that the EU has not harmonised a definition for all Member States to be bound by may lead to further fragmented approaches to content moderation as content in one country may be illegal but not in other Member States. Although there is a definition, there is not much guidance on legal but harmful information, or misinformation, which in the scope of social media platforms may be just as likely to occur through amplification, virality or unintentional interaction.

The difficulty in finding an appropriate definition for all types of false information, and across the EU, has been amplified in the Court decision in *Delfi AS v Estonia* on online hate speech. The Court stressed the difficulties in defining hate speech, and at large, illegal content, in an online environment where numerous considerations are needed.[31] The European Court of Human Rights concluded that hate speech is internationally undefined because there is no universally accepted definition. It encompasses a wide range of messages whether inciting violence or derogatory, based on characteristics like race, colour, religion, or origin.[32] Due to different cultures, languages, and perspectives, it is difficult to define most categories of false information.

[29] Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) (DSA) and Amending Directive 200/31/EC, Article 2(g) and 8(2)(a).
[30] Philippe Juvin and Henna Virkkunen, 'Assessment Of Platforms And Tackling Illegal Content' (*Legislative Train Schedule*, 2022) <https://www.europarl.europa.eu/legislative-train/theme-connected-digital-single-market/file-assessment-of-online-platforms-and-illegal-content> accessed 26 July 2022.
[31] *Delfi v Estonia* [GC] App no. 64569/09 (ECtHR, 16 June 2015).
[32] Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law <http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=URISERV:l33178>.

The issue of defining false information and its sub-categories is further made abundantly clear by Facebook's most recent traumatic shortcomings and its involvement in Myanmar. Facebook's inaction on fake news can be narrowed down to inadequate investment in staff with the necessary linguistic and cultural expertise.[33] Despite admitting to the platform being used to "incite offline violence in Myanmar" through hate speech content, it shows the power Facebook has in disseminating information, despite EU attempts and the need for platforms to be held accountable and do better.[34] Hence, the success of the community standards is close to null. Combatting hate speech requires more than self-regulation and leaves no choice but for government intervention or by other parties.

Facebook's Myanmar shortcomings should not come as any surprise. Facebook is focused only on current issues and wants to quickly sweep them under the carpet and forget about them rather than making the necessary changes to be resilient for future instances of hate speech, misinformation or other issues that may arise.[35] Even suggesting that Facebook needs to alter the wording in their community standards will only go so far in improving their goal of 'transparency' and thus being held accountable for their practices. Nonetheless, Facebook defines hate speech as:

> *"We define hate speech as a direct attack against people – rather than concepts or institutions – on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease."*[36]

Therein, forming tiers of what not to post. This includes expressions using "vile", "disgusting", or "yuck".[37] Facebook, maybe deliberately, uses extremely broad wording, which opens the avenue for many contested cases where expressions that are used do not cause any harm to people. Alternatively, Facebook has only defined misinformation as "content that is false or misleading" and disinformation as "false or misleading posts shared intentionally to deceive

[33] Steve Stecklow, 'Why Facebook is Losing the War on Hate Speech in Myanmar' (*Reuters Investigates,* 2018) <https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/> accessed 15 January 2022.
[34] BBC, 'Facebook admits it was used to 'incite offline violence' in Myanmar' (*BBC,* 2018) <https://www.bbc.co.uk/news/world-asia-46105934> accessed 15 January 2022.
[35] See the Cambridge Analytica scandal.
[36] Facebook (n 8).
[37] Facebook (n 8).

people".[38] Clearer and equally extensive definitions of the types of false information would make for more consistent implementation.

Proposal 1: Rules, definitions, standards, and community guidelines need to be independently developed from platforms to ensure that there is consistent implementation of content moderation practices and adherence to EU definitions.

Despite the difficulties in defining the types of false information, the scope of 'illegal content' should also not be too broad. Broad definitions can be misused against vulnerable groups whom the definitions try to protect, and too narrow definitions can lead to exclusion. If we continue to let Facebook determine what is false information beyond the illegal content categories, we run the risk of allowing social media to limit and tailor our freedom of expression to their needs. Furthermore, Facebook is a private company that is not bound by the same rules as a government and thus making it more difficult to follow up legally as a user. Any new parameters or definitions must continue to ensure that the content that leads to violence or other harm to society is dealt with appropriately in the same way as illegal content is now.

2.1. A Call for Collaboration

Despite difficulties in regulating and defining the types of false information, a collaboration between social media platforms and the EU can ensure harmonisation and a combined and effective effort to minimise the harm false information can have on users.[39]

Proposal 2: Stimulate collaboration with NGOs, social media platforms, civil society, and the EU to create a harmonised definition of illegal content and false information that includes a list of examples and conditions that will satisfy such categorisation. Thereby also setting parameters for legal but harmful content which is more difficult to detect.

[38] Facebook, 'Tacking Action Against Misinformation Across Our Apps' (*Facebook*, 2021) <https://www.facebook.com/combating-misinfo> accessed 18 March 2022.
[39] UNESCO and United Nations Office on Genocide Prevention and the Responsibility to Protect, 'Addressing Hate Speech on Social Media: Contemporary Challenges' (*UNESCO*, 2021) <https://unesdoc.unesco.org/ark:/48223/pf0000379177> accessed 20 January 2022.

Collaborating on a basic set of types of false information that are harmful and can become viral would ensure better protection of users and create a safe online space.[40] An improvement in harmonised definitions and conditions must also consider the shared cultural and political perspectives at a European level to ensure content moderation is not only effective in one jurisdiction but in all Member States, as is the DSA's goal. Without these changes, content moderation practices mean that Facebook and other private platforms will continue to be able to exercise unrivalled power over users' right to freedom of expression and decide what speech is allowed without governmental intervention.

Collaboration between governments and social media platforms must be done in a way that strikes a balance between the interests of both parties and one that does not lead to authoritarian control by governments. Any hard law and enforcement on misinformation and false information at large could further encroach on our human rights, as criticised of the Singaporean approach to implementing a harsh 'fake news law'.[41] Such a law can easily lead to over-blocking content, leaving no room for freedom of expression and undeniably harming democracy.

The balance must ensure that dominant platforms are held accountable for the definitions they set and ensure that Governments do not alone control the definitions that could stifle freedom of speech and human rights. A coherent and harmonised understanding of illegal content and false information will enable platforms in all Member States and jurisdictions to better identify what content must be removed and to do so quickly. Harmonisation will also limit any misunderstanding and grey areas so that illegal and false content cannot slip through the cracks and be viral and amplified in one country but not another.

[40] HM Government, 'Online Harms White Paper' (2019) <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/973939/Online_Harms_White_Paper_V2.pdf> accessed 25 March 2022.

[41] Shibani Mahtani, 'Singapore Introduced Tough Laws Against Fake News. Coronavirus has Put them to the Test' *Washington Post* (Asia and Pacific, 2020) <https://www.washingtonpost.com/world/asia_pacific/exploiting-fake-news-laws-singapore-targets-tech-firms-over-coronavirus-falsehoods/2020/03/16/a49d6aa0-5f8f-11ea-ac50-18701e14e06d_story.html> accessed 18 July 2022.

3. Practical Challenges Concerning Volume and Scope of Content

With an average of 1.97 billion active users on Facebook in 2022,[42] one billion stories shared every day across all Meta's applications in 2018;[43] 300 million monthly active users on Twitter;[44] and 2.1 billion monthly active users on YouTube with more than 500 hours of video uploaded every minute in February of 2020,[45] how can any algorithm or person possibly sift through all the content types accurately and quickly?

To face the immense volume and scope, companies like Facebook spend significant resources on content moderation by employing thousands of moderators worldwide and using sophisticated automation tools to flag and remove content that does not adhere to their community standards.[46] For instance, YouTube and other major platforms determine if pornography is being uploaded by "employing automated text searches for banned words and "skin filters" to determine if a large portion of an image or video shows bare flesh".[47] Zuckerberg has made it clear that it is easier to identify a nipple than to detect misinformation and actual content that can incite violence or harm. However, we still push for more Artificial Intelligence (AI) that cannot pick up on such nuance yet.[48]

Platforms have argued that the policies they have put implemented to moderate content have helped reduce the "aggressiveness and quantity of illegal content online".[49] I argue that this has done little to alleviate the harm of false information and fails to acknowledge that users are innovative and can easily find new platforms to migrate to and continue to grow their following. Especially for those whose content is regularly removed, whether rightfully or not,

[42] Meta, 'Meta Reports Second Quarter 2022 Results' (*Meta*, 2022) <https://s21.q4cdn.com/399680738/files/doc_financials/2022/q2/Meta-06.30.2022-Exhibit-99.1-Final.pdf> accessed 1 August 2022.
[43] Facebook, 'Stories Ad Format' (*Facebook*, 2022) <https://www.facebook.com/business/ads/stories-ad-format#> accessed 1 August 2022.
[44] Mansoor Iqbal, 'Twitter Revenue and Usage Statistics' (*Business of Apps*, 2022) <https://www.businessofapps.com/data/twitter-statistics/> accessed 2 August 2022.
[45] Jack Shepherd, '22 Essential YouTube Statistics You Need to Know in 2022' (*Social Shepherd*, 2022) <https://thesocialshepherd.com/blog/youtube-statistics> accessed 10 August 2022.
[46] Facebook, 'Facebook Community Standards' (*Meta*, 2022) <https://transparency.fb.com/en-gb/policies/community-standards/> accessed 14 March 2022.
[47] Roberts (n 5).
[48] Josh Taylor, 'Not Just Nipples: How Facebook's AI Struggles to Detect Misinformation' (*The Guardian*, 2020) <https://www.theguardian.com/technology/2020/jun/17/not-just-nipples-how-facebooks-ai-struggles-to-detect-misinformation> accessed 19 June 2022.
[49] de Streel (n 28), 43.

their irritations can make it more likely for them to leave mainstream platforms to those that are less rigorously regulated or known, such as Parler or Telegram.[50] However, those like "Parler with inadequate moderation but around 15 million users were denied listing by Apple, Google, and AWS for failing to implement content moderation policies."[51] Sadly, by the time false and harmful information has been posted or users migrate, the harm has already been done. As content moderation currently stands, we are constantly in a tug of war with cyber 'criminals' and false information.

3.1. Community Reporting

Despite the immense number of content uploaded daily, content moderation is inherently inconsistent because it heavily relies on the community reporting, which is not always simple. Although platforms rely on users to report content, actually finding the report button takes some time. It is not as easily accessible as the option to 'like', 'comment' or 'share'. On Facebook it takes 5 clicks before a report is made. Once the user has clicked the 3 dots on the top right corner of a post, the report option is only at the bottom of the drop-down menu and then nudity is the first concern rather than violence or false information.

[50] Sarah Perez, 'Following Riots, Alternative Social Apps and Private Messengers Top the App Stores' (*TechCrunch*, 2021) <https://techcrunch.com/2021/01/11/following-riots-alternative-social-apps-and-private-messengers-top-the-app-stores/> accessed 15 June 2022; Rani Molla, 'Why Right-Wing Extremists' Favorite New Platform is so Dangerous' (*Vox*, 2021) <https://www.vox.com/recode/22238755/telegram-messaging-social-media-extremists> accessed 15 June 2022.

[51] Aggarwal (n 20).

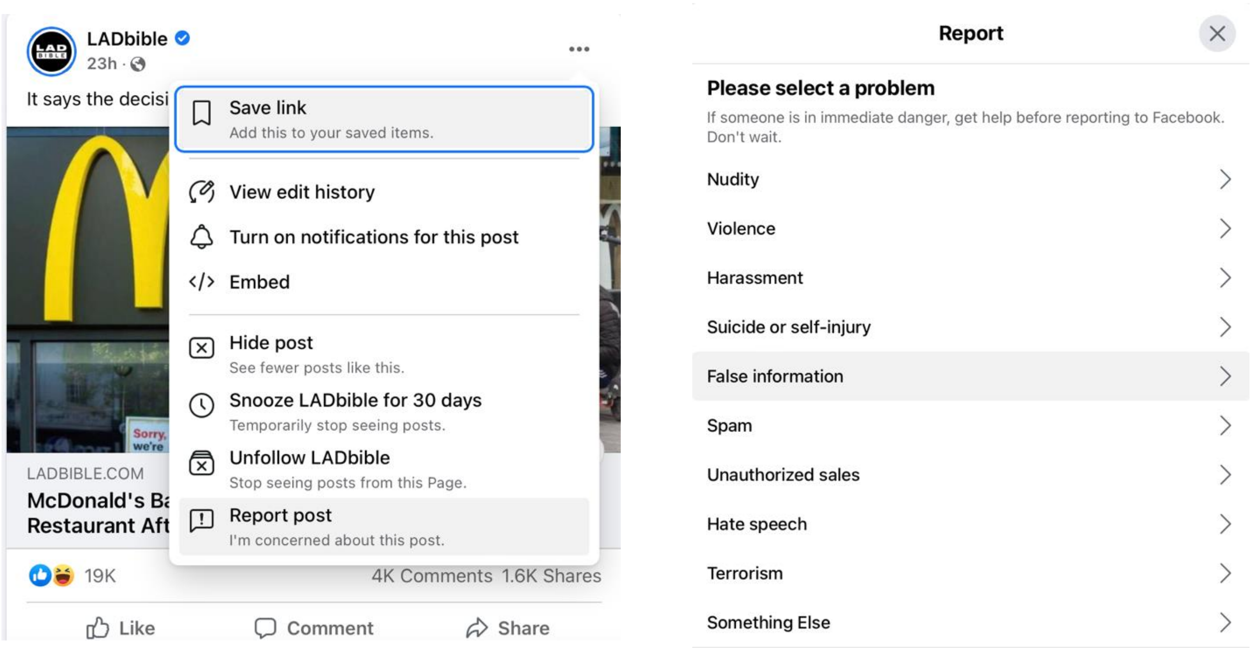
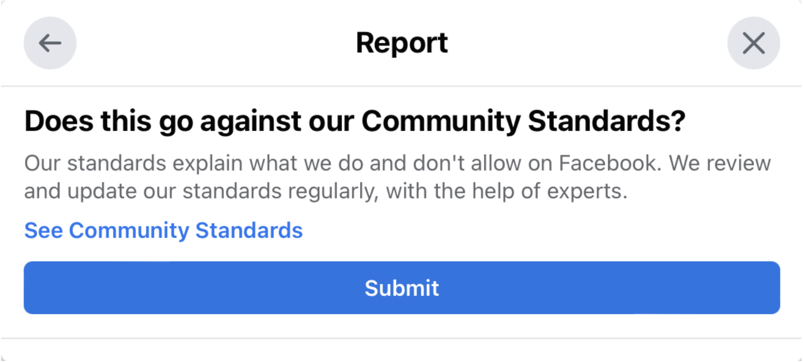


Image 1: Screenshot of Facebook Reporting Options

Reporting content is not as straightforward or accessible to users despite it being the main method for platforms to identify and remove content.

Proposal 3: Platforms must make the report button as clear as the ‘like’, ‘comment’ and ‘share’ options. The report button should be next to where the share button is currently.

Making the report button visible and accessible to users fosters more consistent reporting by users. Simultaneously, although this may lead to more frequent reporting, it may also cause inappropriate reporting. I am weary that in the current format, where the final question before making the report is only whether the content goes against the community standards accompanied by a hyperlink and a submit button.



Does this go against our Community Standards?

Our standards explain what we do and don't allow on Facebook. We review and update our standards regularly, with the help of experts.

See Community Standards

Image 2: Screenshot of Final Step in Reporting Content

Implementing proposal 3 can lead to unjustified complaints or complaints with malicious intent for a specific user. This, in turn, may continue to further burden human moderators and increase the scope of content that needs to be validated.

Proposal 4: Platforms must require users to justify why they want to report content. This extra step will stop users from randomly clicking through and reporting any content. This will also provide better justification and effort on behalf of the flagger.

Unlike using the like button where Facebook can collect data and profile your interactions, no data should be collected on how often a user reports content. This would disincentivize users and likely do more harm to content moderation and users.[52] If platforms rely on users to report content that may be harmful or otherwise aggravating, users should be reminded of their opportunity to do so and make a difference within the digital space. Unfortunately, platforms, with frankly hidden reporting features, have also not established content moderation responsibilities.[53] I am not suggesting a user version of 1984's Big Brother where every user must be on alert and survey the content they do not fully agree with. Rather, users should be informed and regularly reminded of the possibility to flag content and encouraged to question its validity. The same way in which we were taught at school to never trust everything you see online must continue in all our interactions online and on social media platforms. By reminding users of this simple lesson and possibility fosters further protection and user empowerment.

[52] Wu (n 17).
[53] Pariser (n 15).

Proposal 5: Foster an environment where users are reminded of their ability to question the validity of content and help create a safe online space. Users can be reminded through monthly pop-up messages at the top of their feed.

Nevertheless, flagging is the bare minimum as the damage is likely to have already been done and might not be consistently done. Therefore, it is crucial that users are also provided with correct information and from well-informed sources form well rounded opinions.[54] Currently, content flagged as potentially harmful or misleading shows generic banners reminding users to refer to trusted sources like the World Health Organisation (WHO). Essentially users are encouraged to further research content and information that may be false and find the ‘truth’. Self-research could have limited success if users are inclined to confirm their own bias and still rely on social media to find information.[55] Online platforms rely on notice-and-takedown tools to help filter false information but flagging and notice-and-takedown mechanisms are the bare minimum and do not change the fact that the harm has been already done. Further, it lacks any guidelines and can easily lead to over-regulating content that is reported.[56] Flagging is only helpful in so far that it is done consistently and correctly and can stop false information from infiltrating too many users’ newsfeeds.

The costs of entering a platform and sharing content are nil.[57] It is no longer necessary to build a long-term reputation in the same way that is expected of traditional news outlets to be trustworthy. For example, a study on the effect of fake news on the 2016 U.S. Election showed that the “most popular fake news stories were more widely shared on Facebook” than other popular traditional news stories and that those who see fake news were reported to believe them.[58] If it is that easy to persuade and sway users’ opinions on real world and pressing issues,

[54] Adrian Yijie Xu, ‘AI, Truth, and Society: Deepfakes at the front of the Technological Cold War’ (*Medium*, 2019) <https://medium.com/gradientcrescent/ai-truth-and-society-deepfakes-at-the-front-of-the-technological-cold-war-86c3b5103ce6> accessed 22 April 2022.

[55] An Nguyen and Daniel Catalan-Matamoros, 'Digital Mis/Disinformation and Public Engagement with Health and Science Controversies: Fresh Perspectives from Covid-19' (2020) 8(2) Media and Communication 323 <http://doi:10.17645/mac.v8i2.3352> accessed 10 August 2022.

[56] Frederick Mostert and Jane Lambert, ‘Study on IP enforcement measures, especially anti-piracy measures in the digital environment’ (*WIPO Advisory Committee on Enforcement*, 2019) <https://ssrn.com/abstract=3538676> accessed 14 July 2022.

[57] Allcott (n 1), 221.

[58] Chi Luu, ‘The Incredibly True Story of Fake Headlines’ (*JSTOR Daily,* 2019) <https://daily.jstor.org/the-incredibly-true-story-of-fake-headlines/> accessed 4 March 2022.

then it is critical that measures are in place to catch the users' attention by identifying the fake news and providing additional truthful information.[59]

<u>Proposal 6:</u> Verified information must accompany false information so that false information is not amplified unaccompanied and continues to cause potential violence or harm. The verified and correct information should appear as a text alongside the post and be easily comprehendible. Providing additional information allows for a more nuanced view of a topic.

False information should be accompanied by correct and substantiated information. The WHO has played a pivotal role in providing correct information throughout the pandemic and is often referred to as a place to find information on all content referring to COVID-19. On Facebook, the WHO is a verified account which is shown through a blue checkmark. This verification shows the user that the account is who they say they are. Inherently there is an assumption of reliability when afforded to accounts like the WHO.[60] However, these badges are also given to celebrities like Paris Hilton. At times when clarity is of utmost importance and the increase of social media as a news source, I fail to understand why both an agency of the United Nations and an American media personality are awarded the same verification when they have very different roles in the online space and could have opposing views. Following the 2016 Election, where Russian groups were targeting U.S. voters through fake accounts, a verification badge can at least help some users identify abuse from legitimate accounts and sources.[61]

Although the role of a verification badge is to confirm that the account is authentic, the WHO posts important information regarding, say, the pandemic, while celebrities might not. It would be best to differentiate their information from that of celebrities or other verified people to identify fact and fiction. Content posted regarding public health or other noteworthy content should be awarded a banner of correctness rather than only false information being flagged for incorrectness.

[59] Pielemeier (n 11), 924.

[60] Jacqueline Zote, 'Everything You Need To Know About How To Get Verified On Facebook' (*Sprout Blog,* 2020) <https://sproutsocial.com/insights/how-to-get-verified-on-facebook/> accessed 22 July 2022.

[61] Andrew Hutchinson, 'Would Identity Verification Improve Social Media Safety, and Reduce Instances of Trolling and Abuse?' (*Social Media Today,* 2021) <https://www.socialmediatoday.com/news/would-identity-verification-improve-social-media-safety-and-reduce-instanc/596666/> accessed 15 August 2022.

Proposal 7: Accounts like agencies of the United Nations or non-governmental organisations (NGOs) who provide verifiable content should be given a different checkmark to differentiate themselves from accounts such as celebrities, where the blue checkmark only verifies that they are indeed the person they are listed as. The differentiation can be done by using a different colour.

Implementing a different coloured checkmark must only be afforded to accounts that are remarkably engaged with providing correct and verified information for society's benefit, for example public health. The checkmark should not be awarded to politicians, celebrities or other parties who may have their own biases and opinions, as this may lead to further issues of false information and trust in it. I understand that merely providing a different coloured checkmark may not, in the grand scheme of content moderation, make a tremendous difference. However, it may help a few users in distinguishing reliable and fact-checked information from mere opinion and speculation.

Proposal 8: Platforms should also be required to add banners to content that is on their newsfeed based on profiling algorithms. Allowing users to navigate the sea of content and identify what content is from their connections and what is not.

With such an abundance and different types of content, content moderation is no easy task and cannot only depend on users reporting content. The volume of content becomes a two-fold problem. Firstly, that false information becomes harder to detect in the sea of information. Secondly, it means that the information can be digested without questioning its legitimacy or even noticing it and actively absorbing the information.

3.2. Relying on Artificial Intelligence (AI)

To increase the speed and efficiency of checking content and alleviate the workload from human moderators, anything uploaded to Facebook is also reviewed by its automated tools.[62] Although AI can quickly pick up on key words and illegal content, it too can "damage user experience by over-detection and generate false-positives".[63] Without human moderators to

[62] Avaaz, 'Facebook's Algorithm: A Major Threat to Public Health' (*Avaaz,* 2020) <https://secure.avaaz.org/campaign/en/facebook_threat_health/> accessed 29 March 2022.
[63] de Streel (n 28), 44.

pick up on linguistic and cultural nuances, any effort in content moderation is counterproductive. However, picking up on all the nuances is already difficult for the English-speaking community and it becomes even more tricky when different languages and cultures intertwine.

Although AI systems can pick up on basic words like "yuck" or "hate", it cannot differentiate between jokes and sarcasm that is often used on social media platforms like saying, "I hate getting up for school". Looking at more clear-cut examples of illegal content like copyright infringement, AI systems cannot "recognise fair use or other exceptions to the protections of rights holders such as satire or parody".[64]

Proposal 9: Extending proposal 1, platforms must rely on a hyperlocal content moderation approach that collaborates with other languages and contexts to ensure focused and harmonised moderation.

For example, AI cannot always recognise that words in some cultures are offensive but not in others. A classic example being the term 'fag', which in the United Kingdom would mean a cigarette, whereas in the United States, it would be derogatory and homophobic.[65] This is just as difficult for false information, where text classification is much more nuanced and subject to culture, location, and language. Additionally, findings show that false information remains visible without the necessary warnings once the text has been translated from English to another language.[66] What can come across as a harmless phrase when translated to another language could be dangerous in another language or culture. Human moderation still plays a critical role in content moderation, where such expertise is necessary.

3.3. Harms for Human Moderators

Although it may be desirable to rely on human moderators, Roberts has stated that "moderating content from different cultures and regions are tasks that come at high costs (both economic

[64] Amélie Heldt and Stephan Dreyer, 'Competent Third Parties and Content Moderation on Platforms: Potentials of Independent Decision-Making Bodies from a Governance Structure Perspective' (2021) 11(1) Journal of Information Policy 266, <https://www.jstor.org/stable/10.5325/jinfopoli.11.2021.0266> accessed 5 April 2022.
[65] Ruarí Harrison, 'Freedom of Expression and Hate Speech in the EU's Digital Age' (2020) <https://dx.doi.org/10.2139/ssrn.3913882> accessed 10 July 2022.
[66] Avaaz (n 62).

and physiological) when performed by human moderators".[67] Rounds of stories by Facebook's moderators have come to the surface regarding their working conditions. To most, it would be expected that the content moderators must go through would include things like racist jokes or fake news about elections. Unfortunately, the moderators get more than that and see content of "a man having sex with a farm animal [and] graphic videos of a murder".[68] It comes as no surprise that many develop PTSD after having to go through such intense and disturbing content. The working conditions are sadly no better. From being micromanaged to also only some being full-time employees and most being on contract labour, Facebook can keep their high profit margin and get work done around the clock. It was reported that a "content moderator working in Arizona […] will earn just $28,800 per year" compared to an average employee at Facebook who earns $240,000 annually in not only salary, but bonuses and stocks.[69]

<u>Proposal 10:</u> Improve working conditions for human moderators given their role in supporting AI's accuracy, despite the difficulty of content moderation due to the vast volume of content and scope. This must also include properly employing moderators and providing regular mental health checks for all moderators.

With such high expectations of its content moderators, it is highly unfair that the moderators are not treated well nor respected. Such mistreatment can only make it more likely for mistakes in content moderation, and such treatment will not at all improve content moderation. Facebook continues to show its greed in maintaining a high profit, whether through collecting and using users' data or mistreating its moderators.

Content moderation is a tough mechanism to get correct. It can easily lead to the under-blocking of illegal content and false information if the AI systems are given broad definitions and without enough human moderators specialised in different contexts and languages. Likewise, it can also lead to over-blocking if there are too strict requirements and obligations. Both scenarios can jeopardise fundamental freedom of expression rights.

[67] Heldt (n 64), 272.

[68] Casey Newton, 'The Trauma Floor: The Secret Lives of Facebook Moderators in America' (*The Verge*, 2019) <https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona> accessed 25 March 2022; Reuters, 'Ex-Facebook Moderator in Kenya Sues Over Working Conditions' (*The Guardian,* 2022) <https://www.theguardian.com/technology/2022/may/10/ex-facebook-moderator-in-kenya-sues-over-working-conditions> accessed 19 May 2022.

[69] Ibid.

3.3.1. Driving User Protection and Bursting Filter Bubbles

Users are the ones platforms should protect. Instead, they are being used and exploited so that the platforms can maximise their profits.[70] Facebook is not the only one to exploit its users, but like others, Facebook uses models to predict what its users are likely to engage with and be interested in. They collect this information through the user's likes and searches.[71] The algorithm predicts what will be most "valuable and meaningful to an individual over the long term"[72] and creates a web within networks to only show content from those who follow the same or similar interests.[73] Social media platforms divide society through their profiling algorithms, filter bubbles and fake news, all "to keep a public mindlessly clicking and sharing away".[74]

Once you are in a bubble, it is hard to burst it. The algorithm creates tailored and personalised information for the user that they are likely to enjoy but limits their exposure to opinions outside of their own.[75] The lack of diverse opinions and ideas available to a user on their newsfeed may cause "people to develop a distorted view of reality, which may also pave the way for the rapid spread of fake news and rumours".[76] To break these algorithms and filter bubbles, a team of researchers from Finland and Sweden developed an algorithm to increase the diversity of information within bubbles while still allowing content to be shared in the way platforms would like.

In essence, the algorithm reverse engineers the current way in which algorithms ensure virality and amplification. The bubble bursting algorithm provides a diverse feed to a strategic group of users who would then further engage with the information. However, the researchers

[70] Jeff Horwitz and Deepa Seetharaman, 'Facebook Executives Shut Down Efforts to Make the Site Less Divisive' (*The Wall Street Journal*, 2020) <https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499?mod=hp_lead_pos5> accessed 15 April 2022.

[71] Christina Newberry, 'How the Facebook Algorithm Works in 2022 and How to Make it Work for You' (*Hoot Suite*, 2021) <https://blog.hootsuite.com/facebook-algorithm/> accessed 1 June 2022.

[72] Ibid.

[73] Bert-Jaap Koops, 'The Internet and its Opportunities for Cybercrime' (2010) 1(1) Transnational Criminology Manual 735, 3 <http://dx.doi.org/10.2139/ssrn.1738223> accessed 9 May 2022.

[74] Wu (n 17), 322.

[75] Jillian C. York and Corynne McSherry, 'Content Moderation Is Broken. Let Us Count The Ways.' (*Electronic Frontier Foundation*, 2019) <https://www.eff.org/deeplinks/2019/04/content-moderation-broken-let-us-count-ways> accessed 20 June 2022.

[76] Michelle Hampson, 'Smart Algorithm Bursts Social Networks' "Filter Bubbles"' (*IEEE Spectrum*, 2021) <https://spectrum.ieee.org/finally-a-means-for-bursting-social-media-bubbles> accessed 2 August 2022.

recognise that the hurdle is the social media platforms and their willingness to engage with such an algorithm that breaks their maximizing profit goals.

Proposal 11: Platforms must be required to use the new algorithm to diversify users' news feeds so that users are protected, break from confirmation biases, broaden, and challenge their worldview, and make better decisions.

Social media platforms lack a real push to do better; hence, they can hide behind generic banners and flagging. Essentially these gatekeepers become arbiters of truth that can censor speech, amplify speech, and jeopardise basic human rights.[77] Although Facebook has said that they do not want to become the 'arbiters of truth', any AI they use must be transparent to its users and encompass human rights by design.[78] Ideally platforms must either employ the bubble bursting algorithm or otherwise allow users to turn off the recommendation algorithms that tailor content based on their data.

Proposal 12: Develop AI with human-rights concerns by design and appropriately notify users of the algorithms in use. The same algorithm to diversify newsfeeds should be deployed across platforms to allow for a standardised approach to content moderation and create an informed and safe online digital space.

Although platforms, that profit off user data and advertising, should be obliged to implement newsfeed diversification, it is also important that users are aware of the harms of filter bubbles and the ability to be caught in one. Users should be encouraged to follow people on social media with different opinions and to engage in fruitful discussions that may challenge their perspectives.

[77] Nathan Cofnas, 'Deplatforming Won't Work' (*Quillette*, 2019) <https://quillette.com/2019/07/08/deplatforming-wont-work/> accessed 24 April 2022.

[78] Tom McCarthy, 'Zuckerberg says Facebook won't be 'arbiters of truth' after Trump threat' (*The Guardian*, 2020) <https://www.theguardian.com/technology/2020/may/28/zuckerberg-facebook-police-online-speech-trump> accessed 14 March 2022.

4. Significance of Patchwork Regulations

Social media has facilitated the increase of content in comparison to offline outlets. The increase in volume goes through different contexts, cultures, and perspectives, which causes concern for effective global content moderation. The lack of borders in the digital space also makes sharing information across jurisdictions ever easier. Although I am not advocating for digital borders as it may halt innovation and further separate cultures and perspectives. It is important to consider that most dominant platforms, like Facebook, Twitter, and YouTube, originate from the U.S. and their response to content moderation through community standards very much rely on the U.S. legal, political, cultural, and social norms.[79] The American lens also infiltrates through Facebook's use of 'free speech', for which it has adopted an American understanding of protecting the First Amendment and the user's right to free speech. Whereas in the EU, and more widely in international human rights law is broader under the right to freedom of expression and "includes the right to receive information".[80]

Platforms' far-reaching global footprints need to be reeled in and their impact appropriately moderated. Firstly, from a national perspective encompassing national bodies and moderators before tackling the issue of content moderation at a global scale while ensuring that the U.S. perspective is not the norm.

4.1. European Approach to Content Moderation

Dominant social media platforms have a strong global presence but are, at the time of writing, also subject to national laws within the European Union. This means there are several laws that need to be abided by. However, having to comply with several national laws can cause legal uncertainty and chaos instead of succinct content moderation. As a response, the DSA was set to refine and harmonise the moderation process and limit fragmentation within the EU.

To add to the pile of laws, the EU adopted a soft law measure, namely the Code of Practice on Disinformation. Its purpose is guide how tech companies like Facebook, Twitter and YouTube

[79] Langvardt (n 7).

[80] Flynn Coleman, Brandie Nonnecke, and Elizabeth M. Renieris, 'The Promise and Pitfalls of the Facebook Oversight Board' (*Carr Center for Human Rights Policy*, 2021) 4 <https://carrcenter.hks.harvard.edu/files/cchr/files/facebook_oversight_board.pdf> accessed 12 August 2022.

should address disinformation.[81] This includes putting in place safeguards against disinformation, providing transparency on how they ensure accounts devoted to spreading disinformation are limited in their possibilities, such as through deplatforming,[82] invest in technological tools to ensure that authoritative information is displayed whenever appropriate and dilute disinformation's visibility. Deplatforming only works to break social media users connected to specific organisations or people that are devoted to spreading false information. It should also not be relied on as a solution, as users can easily migrate to other platforms to continue to grow their following.[83]

Additionally, the EU has recently accepted a Regulation to address the dissemination of terrorist content online, and the Commission proposed a new legislation to protect against child sexual abuse content and require online platforms to improve their detection mechanisms.[84] Moreover, the European Commission launched a High-Level Expert Group on fake news to further examine the challenge and harm fake news poses to society and to "initiate a reflection on what would be needed at a European Union (EU) level" to protect internet users.[85] Together, the many proposals and the lack of a uniform definition of disinformation, and other types of fake news, have done little to moderate content coherently and continuously let alone provide succinct clarity.

4.2. Digital Services Act replacing the e-Commerce Directive

The DSA, recently approved by the European Parliament, fixes the patchwork of national rules on online harm by creating a singular Act. Previously, Member States, under the e-Commerce Directive, could set their own rules regarding illegal content online. The DSA aims to protect the digital space against illegal content and users' fundamental rights. The Act is novel in attempting to address harmful but lawful content and disinformation. However, there is the

[81] 'Code of Practice on Disinformation' (2018) <https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation> accessed 20 March 2021.

[82] Deplatforming is the action of denying a user access to a platform to express opinions.

[83] Meedan, 'After the Deplatforming: Global Perspectives on Content Moderation' (*Meedan*, 2021) <https://meedan.com/post/after-the-deplatforming-global-perspectives-on-content-moderation> accessed 5 August 2022.

[84] Proposal for a Regulation of The European Parliament and of the Council on preventing the dissemination of terrorist content online; Proposal for a Regulation of The European Parliament and of the Council laying down rules to prevent and combat child sexual abuse.

[85] European Commission, 'Experts Appointed to the High-Level Group on Fake News and Online Disinformation' (*European Commission,* 2018) <*https://digital-strategy.ec.europa.eu/en/news/experts-appointed-high-level-group-fake-news-and-online-disinformation*> accessed 9 July 2022.

critique that establishing an EU wide standard on what is illegal content could be freedom of expression implications if done without the necessary collaboration with relevant stakeholders.[86]

The DSA focuses on relating its rules to existing areas of law including data protection and consumer protection, that already have independent oversight by national regulatory authorities. For example, consumer protection laws are enforced by both national and European bodies.[87] Common to most activities the DSA tries to regulate is data protection. The fear is that with both the General Data Protection Regulation (GDPR) and the DSA regulating the processing of personal data, there may be different outcomes by different authorities. The DSA's biggest shortcoming is, like the GDPR, clarity on enforcement. The DSA indicates that there must be cooperation between national authorities, Boards, and the Commission. However, it fails to give adequate attention to the other authorities and how exactly such cooperation should come about.[88]

The DSA, as soon as next year, will impose several obligations for Very Large Platforms (VLOPs) because of their role in disseminating illegal and harmful content.[89] The obligations include risk assessments, audits and transparency of the algorithms used. It is critical that platforms are transparent with their content moderation practices and prevent false information from becoming viral. When transparency reports form part of a regulatory obligation, social media platforms will likely conform, so public trust in the platform will increase. Transparency is essential to a new regulatory regime especially concerning social media platforms where users go to find information, but the DSA goes further and includes on-site inspections to achieve the necessary transparency.[90] Nevertheless, the requirements for big platforms seem reasonable and maybe even timid.

[86] Joseph Downing, 'The EU's Digital Services Act: Europeanising Social Media Regulation?' (*LSE*, 2022) <https://blogs.lse.ac.uk/europpblog/2022/08/08/the-eus-digital-services-act-europeanising-social-media-regulation/> accessed 13 August 2022.

[87] For example, the European Consumer Organisation.

[88] DSA (n 29), Articles 44 and 46.

[89] Ibid, Article 54.

[90] Ibid; Mark MacCarthy, 'How Online Platform Transparency Can Improve Content Moderation and Algorithmic Performance' (*Brookings*, 2021) <https://www.brookings.edu/blog/techtank/2021/02/17/how-online-platform-transparency-can-improve-content-moderation-and-algorithmic-performance/> accessed 9 July 2022.

Although the obligations are proportionate to the size of the platforms, all platforms must implement a notice-and-takedown mechanism to allow users to notify of illegal content, and all removals must provide a "clear and specific statement of reasons"[91] and users must be able to contest the removal and seek new review.[92] Moreover, platforms that are not established in the EU must appoint a legal representative in the EU.[93]

However, for smaller platforms who might be starting up to rattle the norm, these requirements are expensive, although in theory, the harmonisation will make it easier for smaller platforms to compete with the large ones, to swiftly respond to all notices, publish transparency reports accessibly and rationalise the reasons for take down and acquiring a legal representative is time-consuming and can be costly for small businesses. In addition, non-compliance with these obligations can result in fines of up to 6% of the platform's total turnover.[94] To solve the issue would be to moderate less, which would mean that the platform is not fit for use, and we have a battle again of uncontrollable false information.

Proposal 13: European Bodies must support smaller platforms by teaching affordable and accessible content moderation and implementing safeguards that protect smaller platforms from the costs of the obligations in the DSA.

I strongly advocate for the DSA and its principles as it leaves self-regulation behind, but the biggest struggle with its success is its effect on smaller companies and enforcement. Especially in the long run, as it may become too cumbersome for smaller and start-up platforms to follow. Instead, the DSA could end up sacrificing competition and not offer users the choice of more user friendly, non-exploitative platforms.

The DSA is a respectable tool to safeguard users' human rights through transparency requirements and remedies to content takedown. Nonetheless, it does little to support smaller platforms and needs to ensure that these can enter the market and be given the opportunity to create a safe online space. However, any concern now is all speculation, as only time will tell how the DSA works in practice as of next year.

[91] DSA (n 29), Article 15.
[92] Ibid, Article 17.
[93] Ibid, Article 40(2).
[94] Ibid, Article 59.

5. External Oversight – Friend or Foe

The DSA is proof that any attempt to regulate platforms is laden with complexities. The up till now self-regulation that platforms use leaves a lot of room for them to continue operating without the user's best interests and regulate permitted and prohibited speech. This is naturally a major concern for the freedom of expression.

Although the EU is a trend-setter in the regulatory landscape, the GDPR has shown there are always loopholes that platforms will find, like just bombarding users with consent forms on their websites. Nevertheless, I do not doubt that platforms will continue to find a way around parts of the DSA to continue putting profit and growth before its users without sincere worries about the global problems they contribute to.

Notably, Facebook's approach to self-regulation has only proven that platforms are ill-equipped to deal with such a scale of content and false information. The current legislative frameworks only focus on the responsibility and liabilities of these platforms. I argue for harmonised and transparent external oversight consistent throughout the EU and incorporating fundamental rights from design.

5.1. Oversight Board

During the work on the DSA, Members of the European Parliament proposed to create a "European Social Media Council". Such a council would serve as "an independent advisory group" concerned with "issuing non-binding guiding principles and recommendations to improve content moderation processes, fostering a participative and transparent public debate around content moderation processes; and issuing policy and enforcement recommendations to the commission".[95] This would allow for better cooperation and learning between platforms, external oversight based on international human rights law obligations, better transparency and more diversity in content moderation decisions.

[95] Committee on the Internal Market and Consumer Protection, 'Draft Report with Recommendations to the Commission on Digital Services Act: Improving the Functioning of the Single Market' (European Parliament 2020) <https://www.europarl.europa.eu/doceo/document/IMCO-PR-648474_EN.pdf> accessed 18 June 2022; Article 19, 'Social Media Councils' (*Article 19*, 2021) <https://www.article19.org/wp-content/uploads/2021/10/A19-SMC.pdf> accessed 14 June 2022, 11.

Currently, there is only one attempt at a Social Media Council. Facebook's Oversight Board, hereafter 'the Board', is the external self-regulatory mechanism for the oversight of content moderation. Its first members were installed in May of 2020.[96] As of March 2022, twenty-three decisions have been made.[97] None of which are binding. The founding documents for Facebook's Oversight Board specify that "any prior board decisions will have precedential value and should be viewed as highly persuasive when the facts, applicable policies or other factors are substantially similar".[98] This is troublesome as the goal of the Oversight Board is to "promote free expression by making principled, independent decisions regarding content on Facebook and Instagram".[99] If non-binding, the Board can only offer a small step in building a common ground for content moderation.

Facebook's Oversight Board, on paper, is a display of self-regulation to avoid actual regulatory regulation. Although Facebook has recognised their role in disseminating information and its effects on society, their changes always seem to come too little too late and only when there is a real push from society or regulators to do something.

Facebook's Oversight Board brings the potential for independent assessment of cases that significantly impact human rights. In theory, it can also keep up with technological changes in ways that governments and legislature will always stay a step behind. However, having individual platform-specific Oversight Boards can only lead to more fragmentation and migration. The Oversight Board is mainly involved with addressing high-profile or precedent-setting cases. It should shift its focus on everyday content matters because decisions regarding everyday occurrences will have more impact on users' digestion of average information. Improved self-regulation and external oversight are, without a doubt, a step in the right direction, but it does not go nearly far enough in empowering and protecting users.

[96] Facebook, 'Welcoming the Oversight Board' (*Facebook,* 2020) <https://about.fb.com/news/2020/05/welcoming-the-oversight-board/> accessed 15 January 2022.

[97] Oversight Board, 'Board Decisions' (*Oversight Board,* 2021) <*https*://www.oversightboard.com/decision/> accessed 16 January 2022.

[98] Rory van Loo, 'Federal Rules of Platform Procedure' (2021) 88(4) 867 The University of Chicago Law Review <https://www.jstor.org/stable/27024713> accessed 22 December 2021; Oversight Board, 'Charter' (*Oversight Board,* 2021) <https://about.fb.com/wp-content/uploads/2019/09/oversight_board_charter.pdf /> accessed 15 January 2022, Article 2(2).

[99] Oversight Board, 'The Purpose of the Board' (*Oversight Board,* 2021) <https://oversightboard.com> accessed 15 January 2022.

5.1.1. Members

The Board is a private appeals system which focuses on Facebook and Instagram's content moderation decisions and supports them in what content should be removed or not and its reasoning.[100] The Board is currently composed of 23 members, and there must always be a minimum of 11 and a maximum of 40. Facebook does not name the entire Board but will select a "small group of initial members" where the Board will "lead in selecting all future members" to serve three-year terms.[101] The Board consists of multinational, multidisciplinary individuals whose authority is provided through a trust agreement that explicitly separates the Board from Facebook.[102] Gebhart has rightfully pointed out that there is inadequate representation in the Board "from the Middle East, North Africa, Southeast Asia and missing advocates for LGBTQ+ communities and disability communities".[103] Latonero indicates that the current members undermine the independence and that the decisions will focus entirely on Facebook and its values and policies rather than on human rights standards.[104]

Proposal 14: The Board must follow a specific criterion when choosing its Members to foster diverse perspectives and backgrounds. Representing under-represented regions and communities through the Board's members will ensure a rich assembly of people, backgrounds and expertise and allow for a significant reach.

5.1.2. Independence from Facebook

According to its Charter, the Board is independent of Facebook. The Board's independence should make this type of self-regulation particularly credible and a potential solution to promote free expression. However, the Board being funded by Facebook says otherwise. Facebook gave an initial gift of $130 million to fund the board for the first 6 years.[105] Although

[100] Oversight Board Charter (n 97), Article 1.

[101] Brent Harris, 'Establishing Structure and Governance for an Independent Oversight Board (*Meta*, 2019) <https://about.fb.com/news/2019/09/oversight-board-structure/> accessed 29 March 2022.

[102] Salman (n 21), 29.

[103] Gennie Gebhart, 'How COVID Changed Content Moderation: Year in Review 2020' (*EFF*, 2020) <https://www.eff.org/deeplinks/2020/12/how-covid-changed-content-moderation-year-review-2020> accessed 13 May 2022.

[104] de Streel (n 28).

[105] Kate Klonick, 'The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression' (2020) 129(8) Yale Law Journal 2418, 2469 <https://www.yalelawjournal.org/feature/the-facebook-oversight-board?fbclid=IwAR3ZGhVIeonmFYPa-uyaL4x-qo2XjGNvfzmP2UGEgcXa3Tb4xORdCyjnO9Q> accessed 19 January 2022.

the Board is said to be independent to promote the ‘freedom of speech’, its funding leaves room for inherent bias. Moreover, its independence is further undermined by the rules that govern the appeal process and the Board’s decision-making process.[106] It has also been critiqued that Facebook is still heavily involved and not properly held accountable. This was evident in the most recent decision to suspend Trump’s accounts. Still, the Board stated that in Facebook’s application of a “vague, standardless penalty and then referring this case to the Board to resolve, Facebook seeks to avoid its responsibilities”.[107]

As recommended in the first proposal, by independently creating and enforcing rules and community standards, Facebook cannot hide behind the Oversight Board and will no longer be able to set their own standards for what ‘independence’ means either. By generating genuine independence, there will be less opportunity for a false sense of safety and progress in content moderation.[108] Until then, the independence and place the Board has in practice are not yet crystal clear. Sadly, the Board needs to remind Facebook of its responsibility to protect public safety and respect international human rights.

Proposal 15: Create an Oversight Board that is not funded by one platform. The Board should further engage in independence by involving civil society in the decision-making and indicating which policies have worked and which of the Board’s decisions have affected users and their fundamental rights.

5.1.3. Subject Matter

When Facebook’s moderators have decided to remove a piece of content, the user who produced the content can appeal to it, which means the case is referred to another human moderator for a final decision. The user can then appeal again to the Oversight Board also.

The Board can choose the cases to judge, usually choosing cases which it believes pose precedence or are major issues in content moderation such as censorship or hate speech or

[106] Ibid; Jason Koebler and Joseph Cox, ‘The Impossible Job: Inside Facebook’s Struggle to Moderate Two Billion People’ (*Vice*, 2018) <https://www.vice.com/en/article/xwk9zd/how-facebook-content-moderation-works> accessed 24 June 2022.

[107] Lauren Feiner, ‘Oversight Board Members Defend Punting Decision on Trumps Ban Back to Facebook’ (CNBC, 2021) <https://www.cnbc.com/2021/05/06/oversight-board-members-defend-punting-decision-on-trumps-ban-back-to-facebook.html> accessed 2 August 2022.

[108] Coleman (n 65).

female nudity.[109] The Oversight Board decides which cases to judge based on whether it relates to the Community Standards, their values, or Human Rights Standards.[110] On the surface, it seems that these are justified reasons or parameters, but on the larger scale of content moderation, it means that the Oversight Board is only occupied with a rather narrow scale of content that does not necessarily impact the daily moderation instances.

Proposal 16: Now, a few years in, the Board should expand its scope and review a mix of cases that impact human rights harms, including Facebook's algorithms and amplification on users' news feeds because this concerns more daily encounters with false information.

Facebook's amplification and virality algorithms that are used to retain users can result in adverse impacts on its users. The Oversight Board, in its capacity as an independent Board dedicated to human rights, has the possibility of deciding on those affects and producing a decision on the harms.[111] The Board could set a precedence for all platforms engaging in such amplification algorithms.

5.1.4. Transparency

The Charter, alongside the DSA, includes important transparency provisions. The Board publicises its Members of the Board and each decision will be published online. Facebook publishes quarterly reports on its progress with the non-binding decisions the Board produces. This is in line with the DSA which makes the publication of reports mandatory for providers of intermediary services on any content moderation they have engaged in during that period.[112] This requirement provides stakeholders with a relevant overview of the practices undertaken and the number of complaints Facebook engaged with. However, the Board's Charter does not require the publication of such reports. as the Board directly engages with content moderation and appeals, it is critical that the Board too publishes reports beyond the decisions outlining the number and type of cases it engages with and why.

[109] Elena Debré, 'The Independent Facebook Oversight Board Has Made its First Rulings' (*Slate*, 2021) <https://slate.com/technology/2021/01/facebook-oversight-boards-content-moderation-rulings.html> accessed 25 March 2022.
[110] Oversight Board Charter (n 97), Article 2(2).
[111] BSR, 'Human Rights Review: Facebook Oversight Board' (BSR, 2019) <https://www.bsr.org/reports/BSR_Facebook_Oversight_Board.pdf> accessed 17 July 2022.
[112] DSA (n 29), Article 13.

Proposal 17: Require the Oversight Board to publish bi-annual reports outlining the successes and limitations of the Oversight Board, the number of cases it has engaged with, and the types of content. This will increase the transparency of the Board's functioning and provide an overview for other platforms engaging in content moderation.

Platforms must provide clear information to users about the platforms' decision-making structures and procedures. For instance, Facebook's filtering tools are not entirely public knowledge, which is precisely part of the content moderation and transparency problem.[113] Alongside the reports, the tools platforms use should be simple and easy to understand but also accessible.

5.1.5. Jurisdiction

Although local laws have global effects, better collaboration with social media platforms, will allow Government regulations to be tailored to suit those platforms' capabilities. This can result in more careful content moderation practices across platforms.

However, van Loo rightfully questions whether different platforms' decisions should influence another? For example, should Twitter's resolution of an identical content moderation question but with a different business model to Facebook have any bearing on how Facebook decides a case?[114] I would argue that to create a globally safe digital environment, precedence will need to be relied on, and commonalities between the platforms addressed so that content moderation can be standardised and applicable to platforms in the same way.

I argue that a language analysis of different platforms' community standards and conditions across other platforms and countries need to be analysed for similarities and differences and how to ensure that the gaps are appropriately acknowledged and acted upon.

Proposal 18: Analyse and understand the different community standards of large social media platforms to identify the similarities and differences and create a new community standard guideline that holds for all major platforms. This will ensure that platforms

[113] Salman (n 21), 28.
[114] van Loo (n 98).

operating with similar goals or mechanisms like profiling are moderated similarly, thus improving user safety across platforms.

These recommendations must be considered alongside developments in the DSA to best protect not just users' freedom of expression but also user safety. Facebook's core value is to create a place where users can their views and ideas. However, it fails to acknowledge the abuse the Internet inherently carries across cultures and languages. Content uploads are fast and in large quantities and content moderation needs to keep up without being overshadowed by a dominant platform.

Despite the critique, Facebook offers a new approach to content moderation with external assessment putting international human rights law and freedom of expression first and holding Facebook accountable. Other platforms would also benefit from moving toward greater consistency in the application of their internal policies. Such precedent adds predictability and fairness to content moderation practices.

5.2. The Future of the Oversight Board

The idea behind the Oversight Board has merit. Facebook has even accepted the push from the Board to be more transparent and "develop guidelines for satire, updated nudity detection to protect health-related posts" and has begun efforts against hate speech in other languages.[115]

There remain criticism and recommendations. Article 19 and this report have predominantly called for better transparency, especially within the Oversight Board.[116] External oversight may be a tool best equipped to ensure that platforms are held accountable and that the digital space becomes a safe place for the freedom of expression. I do not oppose the idea of an Oversight Board; frankly, I support that Facebook has taken the initiative given its role and place in the market to improve content moderation. However, this is not without its faults. The Oversight Board is arguably still in its early stages, and it still needs to be seen whether the Board can

[115] Andrew Morse and Queenie Wong 'Facebook says it can't keep up with oversight board's recommendations' (*CNET*, 2021) <https://www.cnet.com/culture/internet/oversight-board-says-facebook-wasnt-forthcoming-on-cross-check-program/> accessed 12 August 2022.

[116] Article 19, 'The Facebook Oversight Board: A Significant step for Facebook and a Small Step for Freedom of Expression' (*Article 19*, 2020) <https://www.article19.org/resources/facebook-oversight-board-freedom-of-expression/> accessed 13 May 2022.

continue to be transparent and improve its independence after the 6 years. It is also still unknown whether Facebook consistently apply the decisions and rules. Nevertheless, I do not doubt that the Oversight Board can set a precedence for other platforms and can make the necessary push to ensure that content moderation is transparent and consistent.

However, based on its current makeup, it starts to feel questionable whether the Oversight Board can be consistent and achieve its goals. Particularly because the Board does not address the underlying problem. By remaining at a platform level, it only affects a small percentage of the larger problem. Content moderation is extremely difficult to hit the right balance consistently. As with any system of moderation, mistakes are bound to happen. While the ability to appeal is an important mechanism to mitigate harm, it does not ensure that fair policies are upheld and even in place – especially given Facebook's overarching role in the Board.

Especially given the global footprint Facebook has, it is worrying whether the Oversight Board, given its current members, will not understand the local contexts and their inherent complexity. Ghosh claims that content moderation's existing issues do not relate to poor moderation practices but to the very core of platforms and their business models.[117] I concur with Ghosh as large platforms like Facebook have continued to show that they would rather focus on user engagement, advertising, harvesting personal data and generating a profit. They are using their consumers rather than serving them, and their overdue responses to social and political turmoil do not suggest otherwise. Changing the business model to truly protect the consumer would be ideal, but unfortunately, it seems this would only be possible if platforms stop in their tracks and build their platforms up again from scratch.

Despite ample critique, Facebook is currently the only social media platform with an external oversight board. I applaud Facebook for presenting a new approach to content moderation with external assessment putting international human rights law and freedom of expression first and holding Facebook accountable. Other platforms would also benefit from moving toward greater consistency in the application of their internal policies. Such precedent adds predictability and fairness to content moderation practices.

[117] Dipayan Ghosh, 'Facebook's Oversight Board is Not Enough' (*Harvard Business Review*, 2019) <https://hbr.org/2019/10/facebooks-oversight-board-is-not-enough> accessed 21 March 2022.

5.3. Expanding Social Media Councils beyond Facebook

Khan, the UN special rapporteur on freedom of expression, has recommended that companies explore the creation of external oversight models such as a Social Media Council (SMC), however, not much has been done to foster this push.[118]

Article 19 and this report's proposal 19 provide the missing push and_propose establishing a SMC to provide an open, transparent, participatory, independent, and accountable forum to review moderation practices.[119] It would be best to start at the European Level. As we have seen with the introduction of the General Data Protection Regulation of 2018, many nations have followed suit and implemented largely the same rules and requirements.[120]

I would advocate for a global framework right away, but that may pose more harm to the freedom of expression and increase the chance of the strictest rules being adopted. Instead, at a national level, the SMC must adhere to international standards, including more involvement by relevant stakeholders and have moderators in local contexts where different cultures, jurisdictions and perspectives can be considered and represented.

Proposal 19: Implement an inter-platform external content moderation body that is entirely devoted to content moderation and is not dependent on one company or a government. This will ensure that false information and content moderation are largely standardised where applicable and not based on a single platform's community standard.

I urge for better collaboration between the legal frameworks and dominant platforms because even though I would not want private companies to become the sole arbiters of truth, they also should not be governments.[121] We must push for a better distribution of power and to reach content moderation decisions that can hold across platforms and with human rights by

[118] Graduate Institute, 'Debating the Facebook Oversight Board' (*Albert Hirschman Centre On Democracy*, 2021) <https://www.graduateinstitute.ch/communications/news/debating-facebook-oversight-board> accessed 9 August 2022; Jillian C. York, 'UN Report Sets Forth Strong Recommendations for Companies to Protect Free Expression' (*EFF*, 2018) <https://www.eff.org/deeplinks/2018/06/un-report-sets-forth-strong-recommendations-companies-protect-free-expression> accessed 19 June 2022.

[119] de Streel (n 28), 61; Article 19 (n 95).

[120] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

[121] Interview with Alexander Urbelis, 'Discussion of content Moderation and Due Process' (2022).

design.[122] As AI is already outsourced to third parties and we require inter-platform content moderation, it would be best suited to implement an independent body entirely dedicated to content moderation. This would support the existing methods of content moderation but also maintain a focus on human rights and ensure a balance between private companies and governments.

Proposal 20: Implement an SMC at the European Level with stringent requirements for transparency and independent review. An independent body and framework for moderation of user-generated online content must put human rights at its very core and not favour a platform.

To be independent and not linked to a singular platform, an SMC by the European Union should be concerned with the investigation of online conduct on all major platforms and provide them clarity and guidance. The SMC would be akin to Facebook's Oversight Board but for all social media platforms, especially dominant ones in the market.

This solution is spurred by a similar initiative by Twitter. Launched in 2019, the BlueSky Initiative would fund an independent research group that would develop decentralised standards for social networks that can be relied on and used by many content moderators.[123] Decentralised content moderation will foster the change from data driven platforms to neutral platforms protecting freedom of expression.[124] This principle within the SMC would make sure that content moderation is no longer held by the dominant few but fairly done by those who truly have the user's interests at the forefront.

Proposal 21: The SMC must be open and participatory, made up of various actors with different backgrounds that work at the national level to protect users and freedom of expression. The SMC must comprise of social media companies, media outlets, regulators,

[122] Heldt (n 64).

[123] Adi Robertson, 'Twitter Is Funding Research Into A Decentralised Version Of Its Platform' (*The Verge,* 2019) <https://www.theverge.com/2019/12/11/21010856/twitter-jack-dorsey-bluesky-decentralized-social-network-research-moderation> accessed 18 March 2022; Adi Robertson, 'Twitter's Decentralised, Open-Source Offshoot Just Released its First Code' (*The Verge,* 2022) <https://www.theverge.com/2022/5/4/23057473/twitter-bluesky-adx-release-open-source-decentralized-social-network> accessed 14 April 2022.

[124] Article 19, 'Why Decentralisation of Content Moderation Might be the Best Way to Protect Freedom of Expression Online' (*Article 19,* 2020) <https://www.article19.org/resources/why-decentralisation-of-content-moderation-might-be-the-best-way-to-protect-freedom-of-expression-online/> accessed 14 April 2022.

advertisers, and civil society organisations. This will appropriately diversify the stakeholders and ensure collaboration and communication.

The complexity of online abuse, hate speech, illegal content, and mis- and disinformation means that many perspectives must be considered for open and transparent discussions. Involving major stakeholders from the beginning can ensure a positive implementation of regulatory frameworks for social media platforms.

An SMC is especially critical because false information always has and always will continue to spread between individuals and groups.[125] Social media platforms have become public market spaces where anyone can share an opinion or an idea freely. Its infrastructure and business models have fostered this possibility and are why information can spread like wildfire with little resistance.[126] An SMC can be a solution that can provide a flexible form of regulation that embraces the innovation of social media platforms and their goals, as well as ensuring that the protection of freedom of expression is a core goal of all digital interactions.

5.3.1. Due Process

Social media platforms play a big role in regulating online speech and have obligations towards their users, governments, and society. It has been openly discussed in this report that when moderators at Facebook or other platforms use algorithms to decide on its users' content and whether to remove it or not, they essentially become the arbiters of truth and judges of truth.

Forcing platforms to engage in content moderation is only successful if they engage in appropriate due process. Facebook engages at several levels with content moderation decisions and has developed their own 'Supreme Court', namely the Oversight Board. Despite requirements set out in the DSA and the Board's Charter, platforms' conduct is not always

[125] Stephan Lewandowsky and Sander van der Linden, 'Countering Misinformation and Fake News Through Inoculation and Prebunking' (2021) 32(2) European Review of Social Psychology, 348 <https://doi.org/10.1080/10463283.2021.1876983> accessed 25 June 2022.

[126] Brandy Zadrozny and Ben Collins, 'How Three Conspiracy Theorists Took Q and sparked QAnon' (*NBC,* 2018) <https://www.nbcnews.com/tech/tech-news/how-three-conspiracy-theorists-took-q-sparked-qanon-n900531> accessed 10 August 2022.

transparent. Beyond Facebook, not every platform even offers a clear appeals procedure, nor is there an independent body concerned with content moderation.[127]

To be successful, such gatekeepers and arbiters of truth must apply these principles in a standardised way. Due process ensures that appeal processes and justice are done throughout the content moderation process. If done correctly, applying due process principles increases the chance of compliance with a decision, platform rules and ensures that human rights are maintained and reviewed, even in the digital space.[128]

The AEQUITAS principles best address the issue of algorithmic content moderation and the need for human review. Frederick Mostert has observed that content moderation on the virality and volume of online speech can swiftly lead to issues of over-blocking and under-blocking and the principle of digital due process is a minimum to protect users' freedom of expression, personal data and intellectual property rights.[129] Thus, the principle of digital due process must guide regulation on content moderation in the online space to ensure that users are given the opportunity to be notified of having content taken down or having a post reviewed, especially through algorithmic moderation and given platforms' responsibility on its users.[130]

The principles supplement the already largely accepted Santa Clara and Manila Principles that guide accountability and transparency. The latter also creates standards for due process and guidelines on notifying users about notice-and-takedowns and the ability to review decisions.[131] Together these principles are minimum standards companies like Facebook and other big social media platforms should implement to ensure that procedural fairness in content moderation protects fundamental rights like the freedom of expression. Likewise, such

[127] Thomas Kadri and Kate Klonick, 'Facebook v. Sullivan: Building Constitutional Law for Online Speech' (2019) 93(1) Southern California Law Review 37 <https://southerncalifornialawreview.com/2019/11/01/facebook-v-sullivan-public-figures-and-newsworthiness-in-online-speech-article-by-thomas-e-kadri-kate-klonick/> accessed 14 June 2022.

[128] Frederick Mostert and Alexander Urbelis, 'Your Day in Court: Social Media Needs a System of Due Process' *(Financial* Times, 2021) <https://www.ft.com/content/48c49453-9a8f-4125-85d7-94220497d13c> accessed 15 February 2022.

[129] Frederick Mostert, ''Digital Due Process': A Need for Online Justice' (2020) Journal of Intellectual Property Law and Practice <https://ssrn.com/abstract=3537058> accessed 4 July 2022.

[130] The Digital Scholarship Institute, 'AEQUITAS Principles' <https://aequitas.online/principles/> accessed 13 February 2022.

[131] Santa Clara Principles, <https://santaclaraprinciples.org> accessed 4 August 2022; Manila Principles <https://www.manilaprinciples.org/> accessed 4 August 2022.

standardisation will further improve transparency and accountability, a main goal this report seeks to provide.

Such digital due processes have proven effective and possible through the implementation of a domain name dispute resolution policy by the World Intellectual Property Organisation (WIPO) that deals with cybersquatting globally.[132] Cybersquatting is the bad faith registration of someone else’s trademark in a domain name. This is an issue that, similarly to false information, is faced with a large volume of content.

The Uniform Dispute Resolution Policy is successful because it set up an independent panel that represents all relevant stakeholders. The success of the WIPO’s approach to combatting cybersquatting is largely due to the collaboration but also in providing easily accessible measures. The WIPO is an example of exactly what a successful SMC can achieve if collaborated appropriately. Global collaboration can promote standardised guidelines to minimise the detriment illegal and false information can have on the functioning of society and user safety.

[132] WIPO, ‘WIPO Guide to the Uniform Domain Name Dispute Resolution’ (*WIPO,* 1999) <https://www.wipo.int/amc/en/domains/guide/> accessed 4 August 2022; Frederick Mostert, ‘The Internet: Regulators Struggle to Balance Freedom with Risk’ <https://www.ft.com/content/e49c39e6-967d-11e9-8cfb-30c211dcd229> accessed 4 August 2022.

5.4. Reset Social Media Completely

This report has shown that there is no one perfect catch all solution to content moderation and that a lot of little solutions need to come together to be meaningful. However, so long as false information is not addressed with appropriate legislation and definitions, the governance of false information will remain in the hands of private and commercial platforms.

An Oversight Board has proven to be a partial solution to content moderation, but the effectiveness of content moderation is not based on an Oversight Board or legislation. Rather, it is the business model that prohibits any meaningful change.[133] It goes almost without saying that who you are in real life and who you are online is slightly different. Online you would like to put your best foot forward and tend to show only the highlights. Platforms on the other hand do not put their best foot forward and hold a vast amount of data on their users and their interactions.[134] Platforms have gone from small places to keep in touch with friends and family to a hyper-connected place that incites violence and hateful conduct. The business model of platforms like Facebook is to maximise consumer engagement by curating feeds and targeting users with ads based on their personal data.[135] Even though platforms are saying they are doing their part to protect users and create a safe digital space by self-regulating and following international standards and human rights, they cannot keep to their word so long as they keep treating users as a mere number in the equation to generate profits.

We have seen with the Oversight Board that it wants to address these issues and remove content that can cause harm and can jeopardise public safety and health. We have also seen how Facebook is still controlling the Board and will want to escape liability and accountability. They have not yet found an alternative way to make a profit without treating their users as less than human. The Oversight Board really just starts to look like something Facebook threw at the problem. The Board, in its current form, cannot actually address the issues Facebook has within its business model and highlights Facebook’s reluctance to accept a rigorous regulatory policy.

[133] Ghosh (n 116).

[134] Wu (n 17).

[135] Ibid; Rae Jereza, ‘Corporeal Moderation: Digital Labour as Affective Good’ (2021) 29(4) Social Anthropology 928, 929 <https://doi.org/10.1111/1469-8676.13106> accessed 4 August 2022.

The Oversight Board is only a temporary bandage for a bigger problem. Although I cannot yet provide a tangible recommendation on what social media 2.0 should look like, I do know that resetting and restarting will at least mean that the data about the user's likes and engagement and the algorithmic bias will lead to a clean slate. In an ideal world, this seems a good starting point to restart.

Even if we were to restart social media, we need to continue to demand for transparency from social media platforms. With an independent body, social media counsel, the human rights principles in the Convention for the Protection of Human Rights and Fundamental Freedoms and the Charter of Fundamental Rights of the European Union need to be the dominant underlying thread to ensure the protection of users and the digital space.

Independent oversight is important to ensure compliance with due process principles. Facebook, as discussed, has taken pivotal steps in ensuring that independent oversight is possible, however, it is not yet adequately independent.

6. Conclusion

Facebook profits from virality and engagement, a model that fosters an ecosystem of false information. Content moderation only plays a small role in changing that business model and protecting users. It alone cannot withstand the pressures from users, civil society and governments to create a truly safe digital space without content that can incite violence or harm. The European Commission can regulate content moderation to better mitigate those harms by ensuring that governments and platforms collaborate on consistent standards that are viable for all platforms that host false information.

While this report seems to suggest one single regulatory regime for all types of content online, it would be naïve to assume that one size fits all. I submit that it would be wise to consider disinformation and misinformation separately due to its inherent difference to blatantly illegal content. Additionally, a firstly EU approach could foster a global effort on platform. Regardless, the DSA is an opportunity to regulate well but it should still incorporate much more of the harm disinformation and misinformation pose and tackle it through bubble bursting algorithms and more transparency requirements.

Harmful but lawful content is still very much uncovered by legislation and the DSA, and any reporting mechanisms and oversight board can only do so little to tackle this. Major platforms like Facebook will continue to prioritise profit, and thereby user data, over users' real safety and only react when absolutely necessary by society. The little too late action has still been the case during the recent U.S. elections and the pandemic. At a bare minimum, we advocate for more robust external oversight, harmonised approaches and greater access to vital knowledge on how and where disinformation crops up in the EU and globally. As soon as we have identified and fixed the inconsistencies in definitions, we have a springboard on how to create a safe online space.

None of these proposals will be possible if platforms do not prioritise users and protect instead of exploit. I am firmly of the opinion that platforms need to start again from the beginning. Each platform keeps building and trying to fix issues but, by design, they are not where they need to be in the 21st century. Platforms must redesign their business models and platforms so that user protection and fundamental rights are at the forefront and centre of their services and products. This is especially critical given that social media platforms are borderless and have a global presence subjected to many laws, cultures, and contexts. So, while we do not yet have a concrete guide on how to *best* moderate content, we know what is not working and what needs to be improved.

The common proposal that acts as a thread throughout this report is a call for collaboration across platforms with all relevant stakeholders. Global collaboration must focus on due process, standardisation, and common nuances. Although it has been suggested to start at the national level, this is only a stepping stone to tackling the issue of content moderation and false information globally. Focusing efforts entirely on national issues of content moderation will only segregate society more and be inadequate in the greater scope of content moderation. The Internet and its content are naturally global and borderless and as such will require an international approach to content moderation. For global oversight and content moderation to work and respect users' fundamental rights, consistent and understandable global guidelines must be set through regular collaboration involving all relevant stakeholders like social media platforms, NGOs, society, and regulators. We must disinfect our newsfeeds with the help of human moderators, AI, digital due process principles and external oversight.

Beyond the scope of this report lie many more proposals that could empower users or at least remind users of their time online and the effect of digesting both correct and incorrect information. However, if users are the ones platforms are protecting, then we are beyond empowering users and need regulatory intervention as proposed in this report.

Appendices

Appendix 1: Summary of Recommendations

Proposal 1: Rules, definitions, standards, and community guidelines need to be independently developed from platforms to ensure that there is consistent implementation of content moderation practices and adherence to EU definitions.

Proposal 2: Stimulate collaboration with NGOs, social media platforms, civil society, and the EU to create a harmonised definition of illegal content and false information that includes a list of examples and conditions that will satisfy such categorisation. Thereby also setting parameters for legal but harmful content which is more difficult to detect.

Proposal 3: Platforms must make the report button as clear as the 'like', 'comment' and 'share' options. The report button should be next to where the share button is currently.

Proposal 4: Platforms must require users to justify why they want to report content. This extra step will stop users from randomly clicking through and reporting any content. This will also provide better justification and effort on behalf of the flagger.

Proposal 5: Foster an environment where users are reminded of their ability to question the validity of content and help create a safe online space. Users can be reminded through monthly pop-up messages at the top of their feed.

Proposal 6: Verified information must accompany false information so that false information is not amplified unaccompanied and continues to cause potential violence or harm. The verified and correct information should appear as a text alongside the post and be easily comprehendible. Providing additional information allows for a more nuanced view of a topic.

Proposal 7: Accounts like agencies of the United Nations or non-governmental organisations (NGOs) who provide verifiable content should be given a different checkmark to differentiate themselves from accounts such as celebrities, where the blue checkmark only verifies that they are indeed the person they are listed as. The differentiation can be done by using a different colour.

Proposal 8: Platforms should also be required to add banners to content that is on their newsfeed based on profiling algorithms. Allowing users to navigate the sea of content and identify what content is from their connections and what is not.

Proposal 9: Extending proposal 1, platforms must rely on a hyperlocal content moderation approach that collaborates with other languages and contexts to ensure focused and harmonised moderation.

Proposal 10: Improve working conditions for human moderators given their role in supporting AI's accuracy, despite the difficulty of content moderation due to the vast volume of content and scope. This must also include properly employing moderators and providing regular mental health checks for all moderators.

Proposal 11: Platforms must be required to use the new algorithm to diversify users' news feeds so that users are protected, break from confirmation biases, broaden, and challenge their worldview, and make better decisions.

Proposal 12: Develop AI with human-rights concerns by design and appropriately notify users of the algorithms in use. The same algorithm to diversify newsfeeds should be deployed across platforms to allow for a standardised approach to content moderation and create an informed and safe online digital space.

Proposal 13: European Bodies must support smaller platforms by teaching affordable and accessible content moderation and implementing safeguards that protect smaller platforms from the costs of the obligations in the DSA.

Proposal 14: The Board must follow a specific criterion when choosing its Members to foster diverse perspectives and backgrounds. Representing under-represented regions and communities through the Board's members will ensure a rich assembly of people, backgrounds and expertise and allow for a significant reach.

Proposal 15: Create an Oversight Board that is not funded by one platform. The Board should further engage in independence by involving civil society in the decision-making and indicating which policies have worked and which of the Board's decisions have affected users and their fundamental rights.

Proposal 16: Now, a few years in, the Board should expand its scope and review a mix of cases that impact human rights harms, including Facebook's algorithms and amplification on users' news feeds because this concerns more daily encounters with false information.

Proposal 17: Require the Oversight Board to publish bi-annual reports outlining the successes and limitations of the Oversight Board, the number of cases it has engaged with, and the types of content. This will increase the transparency of the Board's functioning and provide an overview for other platforms engaging in content moderation.

Proposal 18: Analyse and understand the different community standards of large social media platforms to identify the similarities and differences and create a new community standard guideline that holds for all major platforms. This will ensure that platforms operating with similar goals or mechanisms like profiling are moderated similarly, thus improving user safety across platforms.

Proposal 19: Implement an inter-platform external content moderation body that is entirely devoted to content moderation and is not dependent on one company or a government. This will ensure that false information and content moderation are largely standardised where applicable and not based on a single platform's community standard.

Proposal 20: Implement an SMC at the European Level with stringent requirements for transparency and independent review. An independent body and framework for moderation of user-generated online content must put human rights at its very core and not favour a platform.

Proposal 21: The SMC must be open and participatory, made up of various actors with different backgrounds that work at the national level to protect users and freedom of expression. The SMC must comprise of social media companies, media outlets, regulators, advertisers, and civil society organisations. This will appropriately diversify the stakeholders and ensure collaboration and communication.

Bibliography

Table of Legislation

Code of Practice on Disinformation (2018) <https://digital- strategy.ec.europa.eu/en/policies/code-practice-disinformation>

Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law <http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=URISERV:l33178>

European Convention on Human Rights and Fundamental Freedoms (adopted 1950, entered into force 1953) (ECHR)

Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Network Enforcement Act, NetzDG) <https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.pdf>

Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) (DSA) and Amending Directive 200/31/EC

Proposal for a Regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online

Proposal for a Regulation of the European Parliament and of the Council laying down rules to prevent and combat child sexual abuse

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

UN Guiding Principles, 'Guiding Principles on Business and Human Rights' (*United Nations*, 2011) <https://www.ohchr.org/sites/default/files/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf> accessed 17 March 2022

Universal Declaration of Human Rights (adopted 10 December 1948 UNGA Res 217 A(III) (UDHR)

Case Law

Delfi v Estonia [GC] App no. 64569/09 (ECtHR, 16 June 2015)

Zwarte Piet Facebook Oversight Board Case 2021-002-FB-UA (Oversight Board, April 13, 2021) <https://oversightboard.com/decision/FB-S6NRTDAJ/> 25 March 2022

Articles

Allcott H and Gentzkow M, 'Social Media and Fake News in the 2016 Election' (2017) 31(1) Journal of Economic Perspectives 211, <https://web.stanford.edu/~gentzkow/research/fakenews.pdf> accessed 3 April 2022

Bode L and Vraga E, 'See Something, Say Something: Correction of Global Health Misinformation on Social Media' (2018) 33(9) Health Communication 1131 < https://doi.org/10.1080/10410236.2017.1331312> accessed 15 February 2022

Castets-Renard C, 'Algorithmic Content Moderation on Social Media in EU Law: Illusion of Perfect Enforcement' (2020) University of Illionois Journal of Law, Technology and Policy 1 <https://ssrn.com/abstract=3535107> accessed 9 June 2022

Coleman F, Nonnecke B, and Renieris E, 'The Promise and Pitfalls of the Facebook Oversight Board' (*Carr Center for Human Rights Policy*, 2021) 4 <https://carrcenter.hks.harvard.edu/files/cchr/files/facebook_oversight_board.pdf> accessed 12 August 2022

Gorwa R, Binns R and Katzenbach C, ‘Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance’ (2020) 7 (1) Big Data and Society 1 <https://doi.org/10.1177/2053951719897945> accessed 16 May 2022

Harrison R, ‘Freedom of Expression and Hate Speech in the EU’s Digital Age’ (2020) <https://dx.doi.org/10.2139/ssrn.3913882> accessed 10 July 2022

Heldt A and Dreyer S, ‘Competent Third Parties and Content Moderation on Platforms: Potentials of Independent Decision-Making Bodies from a Governance Structure Perspective’ (2021) 11(1) Journal of Information Policy 266, <https://www.jstor.org/stable/10.5325/jinfopoli.11.2021.0266> accessed 5 April 2022

Jereza R, ‘Corporeal Moderation: Digital Labour as Affective Good’ (2021) 29(4) Social Anthropology 928 <https://doi.org/10.1111/1469-8676.13106> accessed 4 August 2022

Kadri T and Klonick K, ‘Facebook v. Sullivan: Building Constitutional Law for Online Speech’ (2019) 93(1) Southern California Law Review 37 <https://southerncalifornialawreview.com/2019/11/01/facebook-v-sullivan-public-figures-and-newsworthiness-in-online-speech-article-by-thomas-e-kadri-kate-klonick/> accessed 14 June 2022

Klonick K, ‘The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression’ (2020) 129(8) Yale Law Journal 2418 <https://www.yalelawjournal.org/feature/the-facebook-oversight-board?fbclid=IwAR3ZGhVIeonmFYPa-uyaL4x-qo2XjGNvfzmP2UGEgcXa3Tb4xORdCyjnO9Q> accessed 19 January 2022

Koops B, ‘The Internet and its Opportunities for Cybercrime’ (2010) 1(1) Transnational Criminology Manual 735 <http://dx.doi.org/10.2139/ssrn.1738223> accessed 9 May 2022

Langvardt K, ‘Regulating Online Content Moderation’ (2017) 106(5) Georgetown Law Journal 1353, <https://www.law.georgetown.edu/georgetown-law-journal/wp-content/uploads/sites/26/2018/07/Regulating-Online-Content-Moderation.pdf> accessed 4 April 2022

Lewandowsky S and van der Linden S, ‘Countering Misinformation and Fake News Through Inoculation and Prebunking’ (2021) 32(2) European Review of Social Psychology, 348 <https://doi.org/10.1080/10463283.2021.1876983> accessed 25 June 2022

Mostert F, ‘'Digital Due Process': A Need for Online Justice’ (2020) Journal of Intellectual Property Law and Practice <https://ssrn.com/abstract=3537058> accessed 4 July 2022

Mostert F and Lambert J, ‘Study on IP enforcement measures, especially anti-piracy measures in the digital environment’ (*WIPO Advisory Committee on Enforcement*, 2019) <https://ssrn.com/abstract=3538676> accessed 14 July 2022

Nguyen A and Catalan-Matamoros D, 'Digital Mis/Disinformation and Public Engagement with Health and Science Controversies: Fresh Perspectives from Covid-19' (2020) 8(2) Media and Communication 323 <http://doi:10.17645/mac.v8i2.3352> accessed 10 August 2022

Pielemeier J, ‘Disentangling Disinformation: What Makes Regulating Disinformation So Difficult?’ (2020) 4(1) Utah Law Review 917 <https://dc.law.utah.edu/ulr/vol2020/iss4/1> accessed 14 March 2022

Sander B, ‘Freedom Of Expression in the Age on Online Platforms:
the Promise and Pitfalls of a Human Rights- Based Approach to Content Moderation’ (2020) 43(4) Fordham International Law Journal 940 <https://ssrn.com/abstract=3434972> accessed 18 March 2022

van Bavel J and others, 'Using Social and Behavioural Science to Support COVID-19 Pandemic Response' (2020) 4(1) Nature Human Behaviour 460 <https://doi.org/10.1038/s41562-020-0884-z> accessed 9 April 2022

van Loo R, ‘Federal Rules of Platform Procedure’ (2021) 88(4) The University of Chicago Law Review <https://www.jstor.org/stable/27024713> accessed 22 December 2021

Welbers K and Opgenhaffen M, 'Social Media Gatekeeping: An Analysis of The Gatekeeping Influence of Newspapers' Public Facebook Pages' (2018) 20(12) New Media & Society 4728, 4370 <https://doi.org/10.1177/1461444818784302>

Books

Roberts S, 'Understanding Content Moderation', *Behind the Screen: Content Moderation in the Shadows of Social Media* (1st edn, Yale University Press 2019)

van Dijck J, Poell T and De Waal M, *The Platform Society. Public Values In A Connective World* (1st edn, Oxford University Press 2018)

Wu T, *The Attention Merchants* (1st edn, Knopf Publishing Group 2016)

Hearing

Cicciline D in the House Hearing, 116 Congress on 'Online Platforms and Market Power, Part 6: Examining the Dominancy of Amazon, Apple, Facebook and Google' (2020) <https://www.govinfo.gov/content/pkg/CHRG-116hhrg41317/html/CHRG-116hhrg41317.htm> accessed 15 January 2022.

Interview

Interview with Alexander Urbelis, 'Discussion of content Moderation and Due Process' (2022)

Websites

Access Now, 'Access Now's Position on the Digital Services Act Package' (*Access Now,* 2020) <https://www.accessnow.org/cms/assets/uploads/2020/10/Access-Nows-Position-on-the-Digital-Services-Act-Package.pdf> accessed 29 March 2022

Aggarwal M, Dasgupta M and Jaiswal A, 'Safeguarding Social Media: How Effective Content Moderation Can Help Clean Up the Internet' (*Everest Group,* 2021) <https://www.everestgrp.com/safeguarding-social-media-how-effective-content-moderation-can-help-clean-up-the-internet-blog.html> accessed 25 June 2022

Article 19, 'Content Moderation and Freedom of Expression: Bridging the Gap between Social Media and Local Civil Society' (*Article 19,* 2022) <https://www.article19.org/wp-content/uploads/2022/06/Summary-report-social-media-for-peace.pdf> accessed 1 August 2022

Article 19, 'Facebook: New oversight board is not sufficient to safeguard freedom of expression online' (*Article 19,* 2019) <https://www.article19.org/resources/facebook-new-oversight-board-is-not-sufficient-to-safeguard-freedom-of-expression-online/> accessed 4 August 2022

Article 19, 'Social Media Councils' (*Article 19,* 2021) <https://www.article19.org/wp-content/uploads/2021/10/A19-SMC.pdf> accessed 14 June 2022

Article 19, 'The Facebook Oversight Board: A Significant step for Facebook and a Small Step for Freedom of Expression' (*Article 19,* 2020) <https://www.article19.org/resources/facebook-oversight-board-freedom-of-expression/> accessed 13 May 2022

Article 19, 'Why Decentralisation of Content Moderation Might be the Best Way to Protect Freedom of Expression Online' (*Article 19,* 2020) <https://www.article19.org/resources/why-decentralisation-of-content-moderation-might-be-the-best-way-to-protect-freedom-of-expression-online/> accessed 14 April 2022

Avaaz, 'Facebook's Algorithm: A Major Threat to Public Health' (*Avaaz,* 2020) <https://secure.avaaz.org/campaign/en/facebook_threat_health/> accessed 29 March 2022

BBC, 'Covid: Huge Protests Across Europe Over New Restrictions' (2021) <https://www.bbc.co.uk/news/world-europe-59363256> accessed 11 March 2022

BBC, ‘Facebook admits it was used to ‘incite offline violence’ in Myanmar’ (*BBC,* 2018) <https://www.bbc.co.uk/news/world-asia-46105934> accessed 15 January 2022

BSR, ‘Human Rights Review: Facebook Oversight Board’ (BSR, 2019) <https://www.bsr.org/reports/BSR_Facebook_Oversight_Board.pdf> accessed 17 July 2022

Cofnas N, ‘Deplatforming Won’t Work’ (*Quillette*, 2019) <https://quillette.com/2019/07/08/deplatforming-wont-work/> accessed 24 April 2022

Committee on the Internal Market and Consumer Protection, ‘Draft Report with Recommendations to the Commission On Digital Services Act: Improving the Functioning of the Single Market' (European Parliament 2020) <https://www.europarl.europa.eu/doceo/document/IMCO-PR-648474_EN.pdf> accessed 18 June 2022

Cooper P, ‘How the Facebook Algorithm Works in 2021 and How to Work with It’ (*Social Media Marketing and Management Dashboard,* 2021) <https://blog.hootsuite.com/facebook-algorithm/> accessed 1 June 2022

Crawford K, 'Stanford Study Examines Fake News and the 2016 Presidential Election | Stanford News' (*Stanford News,* 2017) <https://news.stanford.edu/2017/01/18/stanford-study-examines-fake-news-2016-presidential-election/> accessed 12 January 2022

Criddle C, ‘Facebook Sued Over Cambridge Analytica Data Scandal’ (*BBC,* 2020) <https://www.bbc.co.uk/news/technology-54722362> accessed 18 February 2022

Debré E, ‘The Independent Facebook Oversight Board Has Made its First Rulings’ (*Slate,* 2021) <https://slate.com/technology/2021/01/facebook-oversight-boards-content-moderation-rulings.html> accessed 25 March 2022

de Posson V, ‘National Initiatives Risk Derailing the DSA’ (*Disruptive Competition* Project, 2021) <https://www.project-disco.org/european-union/030821-national-initiatives-risk-derailing-the-dsa/> accessed 10 May 2022

de Streel A et al, ‘Online Platforms' Moderation of Illegal Content Online Law, Practices and Options for Reform’ (*European Parliament Committee on Internal Market and Consumer Protection,* 2020) <https://www.europarl.europa.eu/RegData/etudes/STUD/2020/652718/IPOL_STU(2020)652718_EN.pdf> accessed 24 March 2022

Docquir P, ‘The Social Media Council: Bringing Human Rights Standards to Content Moderation on Social Media’ (*Centre for International Governance Innovation,* 2019) <https://www.cigionline.org/articles/social-media-council-bringing-human-rights-standards-content-moderation-social-media/> accessed 2 May 2022

DOMO, ‘Data Never Sleeps 9.0’ (*DOMO,* 2021) <https://web-assets.domo.com/blog/wp-content/uploads/2021/09/data-never-sleeps-9.0-1200px-1.png> accessed 15 July 2022

Downing J, ‘The EU’s Digital Services Act: Europeanising Social Media Regulation?’ (*LSE,* 2022) <https://blogs.lse.ac.uk/europpblog/2022/08/08/the-eus-digital-services-act-europeanising-social-media-regulation/> accessed 13 August 2022

EDRi, ‘Platform Regulation Done Right’ (*EDRi,* 2020) <https://edri.org/wp-content/uploads/2020/04/DSA_EDRiPositionPaper.pdf> accessed 24 March 2022

European Commission, ‘Experts Appointed to the High-Level Group on Fake News and Online Disinformation’ (*European Commission,* 2018) <*https://digital-strategy.ec.europa.eu/en/news/experts-appointed-high-level-group-fake-news-and-online-disinformation*> accessed 9 July 2022

Facebook, ‘Community Standards, Hate Speech’ (*Facebook,* 2022) <*https*://transparency.fb.com/en-gb/policies/community-standards/hate-speech/> accessed 13 January 2022

Facebook, ‘Facebook Community Standards’ (*Meta,* 2022) <https://transparency.fb.com/en-gb/policies/community-standards/> accessed 14 March 2022

Facebook, ‘How can I manage the Time I Spend on Facebook?’ (*Facebook,* 2022) <https://m.facebook.com/help/1737706169659354/iphone-app-help/?helpref=platform_switcher&cms_platform=iphone-app> accessed 14 August 2022.

Facebook, ‘Stories Ad Format’ (*Facebook*, 2022) <https://www.facebook.com/business/ads/stories-ad-format#> accessed 1 August 2022

Facebook, ‘Tacking Action Against Misinformation Across Our Apps’ (*Facebook*, 2021) <https://www.facebook.com/combating-misinfo> accessed 18 March 2022

Facebook, ‘Welcoming the Oversight Board’ (*Facebook*, 2020) <https://about.fb.com/news/2020/05/welcoming-the-oversight-board/> accessed 15 January 2022

Fazio L, 'Out-Of-Context Photos Are A Powerful Low-Tech Form Of Misinformation' (*The Conversation*, 2020) <https://theconversation.com/out-of-context-photos-are-a-powerful-low-tech-form-of-misinformation-129959> accessed 13 June 2022.

Feiner L, ‘Oversight Board Members Defend Punting Decision on Trumps Ban Back to Facebook’ (CNBC, 2021) <https://www.cnbc.com/2021/05/06/oversight-board-members-defend-punting-decision-on-trumps-ban-back-to-facebook.html> accessed 2 August 2022

Fung B, ‘Facebook’s Oversight Board is Finally Hearing Cases, Two Years After it was First Announced’ (*CNN Business,* 2020) <https://edition.cnn.com/2020/10/22/tech/facebook-oversight-board/index.html> accessed 24 March 2022

Furlan M, ‘The New Code of Practice on Disinformation: An Attempt to Restore Responsibility in the Online Public Sphere’ (*89 Initiative,* 2021) <https://89initiative.com/the-new-code-of-practice-on-disinformation-an-attempt-to-restore-responsibility-in-the-online-public-sphere/> accessed 23 March 2022

Gebhart G, ‘How COVID Changed Content Moderation: Year in Review 2020’ (*EFF*, 2020) <https://www.eff.org/deeplinks/2020/12/how-covid-changed-content-moderation-year-review-2020> accessed 13 May 2022

Ghosh D, ‘Facebook’s Oversight Board is Not Enough’ (*Harvard Business Review,* 2019) <https://hbr.org/2019/10/facebooks-oversight-board-is-not-enough> accessed 21 March 2022

Graduate Institute, ‘Debating the Facebook Oversight Board’ (*Albert Hirschman Centre On Democracy,* 2021) <https://www.graduateinstitute.ch/communications/news/debating-facebook-oversight-board> accessed 9 August 2022

Hampson M, 'Smart Algorithm Bursts Social Networks' "Filter Bubbles"' (*IEEE Spectrum,* 2021) <https://spectrum.ieee.org/finally-a-means-for-bursting-social-media-bubbles> accessed 2 August 2022

Harris B, ‘Establishing Structure and Governance for an Independent Oversight Board (*Meta,* 2019) <https://about.fb.com/news/2019/09/oversight-board-structure/> accessed 29 March 2022

HM Government, ‘Online Harms White Paper’ (2019) <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/973939/Online_Harms_White_Paper_V2.pdf> accessed 25 March 2022

Horwitz J and Seetharaman D, ‘Facebook Executives Shut Down Efforts to Make the Site Less Divisive’ (*The Wall Street Journal,* 2020) <https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499?mod=hp_lead_pos5> accessed 15 April 2022

Hutchinson A, ‘Would Identity Verification Improve Social Media Safety, and Reduce Instances of Trolling and Abuse?’ (*Social Media Today,* 2021) <https://www.socialmediatoday.com/news/would-identity-verification-improve-social-media-safety-and-reduce-instanc/596666/> accessed 15 August 2022

Iqbal M, ‘Twitter Revenue and Usage Statistics’ (*Business of Apps,* 2022) <https://www.businessofapps.com/data/twitter-statistics/> accessed 2 August 2022

Juvin P, and Virkkunen H, 'Assessment Of Platforms And Tackling Illegal Content' (*Legislative Train Schedule,* 2022) <https://www.europarl.europa.eu/legislative-train/theme-connected-digital-single-market/file-assessment-of-online-platforms-and-illegal-content> accessed 26 July 2022

Khan I, ‘Disinformation and Freedom of Opinion and Expression’ (*UN Human Rights Council,* 2021) <https://teaching.globalfreedomofexpression.columbia.edu/node/479> accessed 9 August 2022

Koebler J and Cox J, ‘The Impossible Job: Inside Facebook’s Struggle to Moderate Two Billion People’ (*Vice,* 2018) <https://www.vice.com/en/article/xwk9zd/how-facebook-content-moderation-works> accessed 24 June 2022

Luu C, ‘The Incredibly True Story of Fake Headlines’ (*JSTOR Daily,* 2019) <https://daily.jstor.org/the-incredibly-true-story-of-fake-headlines/> accessed 4 March 2022

MacCarthy M, ‘How Online Platform Transparency Can Improve Content Moderation and Algorithmic Performance’ (*Brookings,* 2021) <https://www.brookings.edu/blog/techtank/2021/02/17/how-online-platform-transparency-can-improve-content-moderation-and-algorithmic-performance/> accessed 9 July 2022.

McCarthy T, ‘Zuckerberg says Facebook won't be 'arbiters of truth' after Trump threat’ (*The Guardian*, 2020) <https://www.theguardian.com/technology/2020/may/28/zuckerberg-facebook-police-online-speech-trump> accessed 14 March 2022

Mahtani S, ‘Singapore Introduced Tough Laws Against Fake News. Coronavirus has Put them to the Test’ *Washington Post* (Asia and Pacific, 2020) <https://www.washingtonpost.com/world/asia_pacific/exploiting-fake-news-laws-singapore-targets-tech-firms- over-coronavirus-falsehoods/2020/03/16/a49d6aa0-5f8f-11ea-ac50-18701e14e06d_story.html> accessed 18 July 2022

Manila Principles <https://www.manilaprinciples.org/> accessed 4 August 2022

Meedan, ‘After the Deplatforming: Global Perspectives on Content Moderation’ (*Meedan,* 2021) <https://meedan.com/post/after-the-deplatforming-global-perspectives-on-content-moderation> accessed 5 August 2022

Meserole M, ‘How Misinformation Spreads on Social Media – and What to do About It’ (*Brookings,* 2018) <https://www.brookings.edu/blog/order-from-chaos/2018/05/09/how-misinformation-spreads-on-social-media-and-what-to-do-about-it/> accessed 14 August 2022

Meta, ‘Meta Reports Second Quarter 2022 Results’ (*Meta*, 2022) <https://s21.q4cdn.com/399680738/files/doc_financials/2022/q2/Meta-06.30.2022-Exhibit-99.1-Final.pdf> accessed 1 August 2022

Molla R, ‘Why Right-Wing Extremists’ Favorite New Platform is so Dangerous’ (*Vox,* 2021) <https://www.vox.com/recode/22238755/telegram-messaging-social-media-extremists> accessed 15 June 2022

Morse A and Wong Q ‘Facebook says it can’t keep up with oversight board’s recommendations’ (*CNET,* 2021) <https://www.cnet.com/culture/internet/oversight-board-says-facebook-wasnt-forthcoming-on-cross-check-program/> accessed 12 August 2022

Mostert F, ‘The Internet: Regulators Struggle to Balance Freedom with Risk’ <https://www.ft.com/content/e49c39e6-967d-11e9-8cfb-30c211dcd229> accessed 4 August 2022

Mostert F and Urbelis A, ‘Your Day in Court: Social Media Needs a System of Due Process’ *(Financial* Times, 2021) <https://www.ft.com/content/48c49453-9a8f-4125-85d7-94220497d13c> accessed 15 February 2022

Newberry C, ‘How the Facebook Algorithm Works in 2022 and How to Make it Work for You’ (*Hoot Suite,* 2021) <https://blog.hootsuite.com/facebook-algorithm/> accessed 1 June 2022

Newton C, ‘The Trauma Floor: The Secret Lives of Facebook Moderators in America’ (*The Verge*, 2019) <https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona> accessed 25 March 2022

Ohrvik-Stott J and Miller C, ‘A Digital Duty of Care’ (*Doteveryone,* 2019) <https://doteveryone.org.uk/wp-content/uploads/2019/02/Doteveryone-briefing-a-digital-duty-of-care.pdf> accessed 28 February 2022

O’Kane S, ‘New Study will Show Misinformation on Facebook gets way more Engagement than News’ (*The Verge,* 2021) <https://www.theverge.com/2021/9/3/22656036/nyu-researchers-study-facebook-misinformation-engagement-election> accessed 28 March 2022

Oversight Board, ‘Board Decisions’ (*Oversight Board,* 2021) <*https*://www.oversightboard.com/decision/> accessed 16 January 2022

Oversight Board, ‘Charter’ (*Oversight Board,* 2021) <https://about.fb.com/wp-content/uploads/2019/09/oversight_board_charter.pdf> accessed 15 January 2022

Oversight Board, ‘The Purpose of the Board’ (*Oversight Board,* 2021) <https://oversightboard.com> accessed 15 January 2022

Perez S, ‘Following Riots, Alternative Social Apps and Private Messengers Top the App Stores’ (*TechCrunch,* 2021) <https://techcrunch.com/2021/01/11/following-riots-alternative-social-apps-and-private-messengers-top-the-app-stores/> accessed 15 June 2022

Rastogi P, ‘Should Social Media Platforms be the Arbitrator of Truth?’ (*Medium,* 2021) <https://medium.com/redhill-review/should-social-media-platforms-be-the-arbitrator-of-truth-760df94baa70> accessed 13 August 2022

Reuters, ‘Ex-Facebook Moderator in Kenya Sues Over Working Conditions’ (*The Guardian,* 2022) <https://www.theguardian.com/technology/2022/may/10/ex-facebook-moderator-in-kenya-sues-over-working-conditions> accessed 19 May 2022

Robertson A, ‘Twitter’s Decentralised, Open-Source Offshoot Just Released its First Code’ (*The Verge,* 2022) <https://www.theverge.com/2022/5/4/23057473/twitter-bluesky-adx-release-open-source-decentralized-social-network> accessed 14 April 2022

Robertson A, ‘Twitter Is Funding Research Into A Decentralised Version Of Its Platform’ (*The Verge,* 2019) <https://www.theverge.com/2019/12/11/21010856/twitter-jack-dorsey-bluesky-decentralized-social-network-research-moderation> accessed 18 March 2022

Salman H, ‘Regulating the Digital Resonance’ (*American University Washington College of Law,* 2021) <https://digitalcommons.wcl.american.edu/cgi/viewcontent.cgi?article=1053&context=stu_upperlevel_papers> accessed 21 March 2022

Santa Clara Principles, <https://santaclaraprinciples.org> accessed 4 August 2022

Satariano A, ‘E.U. Takes Aim at Social Media’s Harms with Landmark New Law’ (*The New York Times,* 2022) <https://www-nytimes-com.cdn.ampproject.org/c/s/www.nytimes.com/2022/04/22/technology/european-union-social-media-law.amp.html> accessed 29 May 2022

Shepherd J, ’22 Essential YouTube Statistics You Need to Know in 2022’ (*Social Shepherd,* 2022) <https://thesocialshepherd.com/blog/youtube-statistics> accessed 10 August 2022

Sosa A and others, 'An Open Letter to Spotify' (*An Open Letter to Spotify*, 2021) <https://spotifyopenletter.wordpress.com/2022/01/10/an-open-letter-to-spotify/> accessed 17 January 2022

Stecklow S, ‘Why Facebook is Losing the War on Hate Speech in Myanmar’ (*Reuters Investigates,* 2018) <https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/> accessed 15 January 2022

Taylor J, ‘Not Just Nipples: How Facebook’s AI Struggles to Detect Misinformation’ (*The Guardian*, 2020) <https://www.theguardian.com/technology/2020/jun/17/not-just-nipples-how-facebooks-ai-struggles-to-detect-misinformation> accessed 19 June 2022

The Digital Scholarship Institute, ‘AEQUITAS Principles’ <https://aequitas.online/principles/> accessed 13 February 2022

UNESCO and United Nations Office on Genocide Prevention and the Responsibility to Protect, 'Addressing Hate Speech on Social Media: Contemporary Challenges' (*UNESCO*, 2021) <https://unesdoc.unesco.org/ark:/48223/pf0000379177> accessed 20 January 2022

United Nations Human Rights Office of the High Commissioner, ‘Moderating Online Content: Fighting Harm or Silencing Dissent?’ (OHCHR, 2021) <https://www.ohchr.org/en/stories/2021/07/moderating-online-content-fighting-harm-or-silencing-dissent> accessed 19 July 2022

Waterson J and Milmo D, ‘Facebook Whistleblower Frances Haugen Calls for Urgent External Regulation’ (*The Guardian*, 2021) <https://www.theguardian.com/technology/2021/oct/25/facebook-whistleblower-frances-haugen-calls-for-urgent-external-regulation> accessed 19 July 2022

WIPO, ‘WIPO Guide to the Uniform Domain Name Dispute Resolution’ (*WIPO*, 1999) <https://www.wipo.int/amc/en/domains/guide/> accessed 4 August 2022

Xu A, ‘AI, Truth, and Society: Deepfakes at the front of the Technological Cold War’ (*Medium*, 2019) <https://medium.com/gradientcrescent/ai-truth-and-society-deepfakes-at-the-front-of-the-technological-cold-war-86c3b5103ce6> accessed 22 April 2022

York J and McSherry C, 'Content Moderation Is Broken. Let Us Count The Ways.' (*Electronic Frontier Foundation*, 2019) <https://www.eff.org/deeplinks/2019/04/content-moderation-broken-let-us-count-ways> accessed 20 June 2022

York J, ‘UN Report Sets Forth Strong Recommendations for Companies to Protect Free Expression’ (*EFF*, 2018) <https://www.eff.org/deeplinks/2018/06/un-report-sets-forth-strong-recommendations-companies-protect-free-expression> accessed 19 June 2022.

York J, ‘Social media companies say they want to be transparent. So why aren’t they?’ (*Wired*, 2021) <https://www.wired.co.uk/article/santa-clara-principles> accessed 1 August 2022.

Zadrozny B and Collins B, ‘How Three Conspiracy Theorists Took Q and sparked QAnon’ (*NBC*, 2018) <https://www.nbcnews.com/tech/tech-news/how-three-conspiracy-theorists-took-q-sparked-qanon-n900531> accessed 10 August 2022

Zote J, 'Everything You Need To Know About How To Get Verified On Facebook' (*Sprout Blog*, 2020) <https://sproutsocial.com/insights/how-to-get-verified-on-facebook/> accessed 22 July 2022

Videos

Pariser E, 'Beware Online "Filter Bubbles"' <https://www.ted.com/talks/eli_pariser_beware_online_filter_bubbles> accessed 17 July 2022