

https://dspace.org/

Social media content moderation and limitations on freedom of expression: the role of the state

Davis, Tara

2022-10-30
University of the Witwatersrand, Johannesburg
https://hdl.handle.net/10539/38598
Davis, Tara. (2024). The role of design houses [Master's dissertation PhD thesis, University of the Witwatersrand, Johannesburg]. Wiredspace. https://hdl.handle.net/10539/38598

Downloaded from a DSpace repository. https://dspace.org/

SOCIAL MEDIA CONTENT MODERATION AND LIMITATIONS ON FREEDOM OF EXPRESSION: THE ROLE OF THE STATE

by

Tara Davis

Supervisors: Franziska Sucker an Mariya Badeva-Bright

Submitted in partial fulfilment of the requirements for the degree of

Master of Laws by Coursework and Research Report

at the University of the Witwatersrand, Johannesburg

Date:30 October 2022

Abstract

Social media platforms are an important space for the exercise of freedom of expression. However, their operation poses a concomitant risk to the right. The right is both protected by and undermined by content moderation - a mechanism that enables platforms to determine the bounds of permissible speech, and what expression to highlight or suppress. This has significant implications for the right to freedom of expression but has been left largely to the private sector to regulate. This raises important questions about the role of the state in protecting and enabling the right to freedom of expression online. In this paper I discuss the need for content moderation, how it operates and the ways in which it undermines the right to freedom of expression in South Africa. I further explore the role of the state and analyse South Africa's current regulatory response. In so doing, I contend that South Africa's approach does not adequately respect, protect and promote the right to freedom of expression.

Contents

I. INTRODUCTION 4

II. THE UNREALISED PROMISE OF THE INTERNET AND WHY WE MODERATE 5

(a) The need for content moderation 5

(b) The rationale for moderation and how it is defined 7

(c) The operation of content moderation 8

(d) Considerations for content moderation 12

III. THE RIGHT TO FREEDOM OF EXPRESSION IN SOUTH AFRICA 14

(a) Un-packing the right to freedom of expression in South Africa 14

(b) The operation and limitations of the right to freedom of expression 15

(c) Freedom of expression online 17

IV. HOW CONTENT MODERATION UNDERMINES THE RIGHT TO FREEDOM OF EXPRESSION IN SOUTH AFRICA 18

(a) The limitations extend beyond section 16(2) 19

(b) The limitations are vague 19

(c) Prior restraint 20

V. THE ROLE OF THE STATE 22

(a) Why involve the state? 22

(b) Reconceptualising the right in South Africa 24

VI. SOUTH AFRICA'S REGULATION OF ONLINE EXPRESSION 26

(a) Conditional limitation of liability 26

(b) The take-down notification process 27

(i) Incentivising removal 28

(ii) Outsourcing rights-based determinations 28

(c) The complaints mechanism 29

(d) The obligation to classify 30

VII. CONCLUSION 31

I. INTRODUCTION

In 2020, three black footballers received hundreds of racially abusive comments on their social media profiles,[1] many of which included an emoji of a monkey. Despite multiple reports of abuse, a significant amount of the content remained online, despite its harmful, racist nature.[2] In the same year, activists from Nigeria who were protesting police brutality and the Nigerian government, had their content blocked. Despite the public importance of disseminating the information, the content was incorrectly flagged as false and removed from social media platforms.[3]

These examples demonstrate the need for content moderation online and its potential harm. Abusive content – such as hate speech– may infringe on the rights of others and undermine their safety online and offline. However, incorrect removal may amount to censorship and infringe the right to freedom of expression and other associated rights. There has accordingly been significant debate around the appropriate balance between protecting freedom of expression and mitigating against harm online. This balance is mediated through content moderation – the process through which platforms determine the bounds and operation of permissible content on their site. It is no easy task and one which has been left largely to the private sector.

What these examples also allude to is the context within which content moderation occurs. A context where moderation choices are dictated by power and influence, and sometimes political interference. This context, coupled with the significant implications that content moderation has on the right to freedom of expression raises questions about the appropriate role of the state in the moderation process.

In this paper, I discuss the role of the state in limiting the implications of content moderation on social media platforms on the right to freedom of expression in South Africa. In so doing, I focus

[1] Victoria Elms & Kieran Devine 'Euro 2020: Why is it so difficult to track down racist trolls and remove hateful messages on social media?' *Sky News* 21 July 2021 at 1, available at *https://news.sky.com/story/euro-2020-why-is-it-so-difficult-to-track-down-racist-trolls-and-remove-hateful-messages-on-social-media-12358392*, accessed on 4 February 2022.

[2] David Wicock 'Instagram claims monkey emojis "DON'T breach race rules" as Twitter removes 1,000 racist posts after appalling abuse of England stars' *Daily Mail* 12 July 2021 at 1, available at *https://www.dailymail.co.uk/news/article-9780627/Racist-tweeters-knock-door-police-face-force-law.html*, accessed on 4 February 2022.

[3] David Gilbert 'Facebook and Instagram are Censoring Protests Against Police Violence in Nigeria' *Vice* 21 October 2021 at 2, available at *https://www.vice.com/en/article/jgqeyg/facebook-is-censoring-protests-against-police-violence-in-nigeria,* accessed on 4 February 2022.

on five things. First, content moderation – why we need it, what it is and how it works. Second, the right to freedom of expression in South Africa – why it matters, how it may be lawfully limited, and how it operates online. Third, the implications that content moderation has on the right to freedom of expression – including how it limits the right. Fourth, the role of the state and finally, South Africa's current regulation – what it's doing, and whether it's sufficient.

These focus areas culminate in a conclusion that the state's regulatory measures inadequately extinguish its duty to respect and protect the right to freedom of expression.

II. THE UNREALISED PROMISE OF THE INTERNET AND WHY WE MODERATE

(a)The need for content moderation

The internet has fundamentally changed the exercise of the right to freedom of expression. Its revolutionary potential was attributed to its participatory nature as an interactive medium.[4] Unlike other communication technology, such as print and television, internet users are not confined to the role of passive observer but can actively engage as creators and publishers of their own content.[5] The internet promised to elevate anyone's content and remove the publisher as gatekeeper.[6] Further, engagement on the internet is inexpensive, does not require specialist equipment, provides a large, global audience, and may provide a degree of anonymity.[7] These characteristics posed an opportunity for unprecedented, unconstrained freedom of expression, making early users of the internet hopeful of its potential as an open, participatory and democratic space.[8]

Years and experience, however, proved that early hope to be a fallacy. Instead, digital spaces were overrun by harmful content and conduct including things such as spam, violence, hate speech, abuse and harassment. According to Grimmelman, 'that's just how people act on the Internet.'[9]

---

[4] Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression Doc A/HRC/17/27 16 May 2011 at 6.
[5] Ibid.
[6] Tarleton Gillespie *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media* (2018) 16.
[7] Fawzia Cassim 'Regulating hate speech and freedom of expression on the internet: Promoting tolerance and diversity' (2015) 28(3) *South African Journal of Criminal Justice* 303 at 305.
[8] Gillespie op cit note 6 at 16.
[9] James Grimmelmann 'The virtues of moderation' (2015) 17 *Yale Journal of Law and Technology* 42 at 45.

Despite his rather glib remark, online abuse is a pervasive and significant problem that threatens the safety, well-being, and rights of users.

Harmful content is prevalent online. In just 3 months Facebook acted on 13.5 million pieces of content it considered hate speech,[10] 45.9 pieces of violent and graphic content and[11] 8.2 million pieces identified as harassment.[12] Some content is targeted at a specific user – where expression is used to attack a user's gender, sexual orientation or race.[13] This is a form of harassment, which is disproportionality experienced by women and racial minorities.[14] Some content is disseminated with the intent to harm a broader community, referred to as trolling.[15] In these instances a user intentionally posts antagonistic information to a forum or group to elicit a hostile response.[16] In its most extreme form, some expression entails an orchestrated and intentional attack by an entire community aimed at harming society.[17]

Abusive content is not the only form of harmful content – the dissemination of false information poses a challenge too. This is not a new phenomenon, but the internet has proliferated the concern.[18] Now, anyone can develop and share false information.[19] Research suggests that people prefer to access information shared on social media,[20] and it spreads faster than correct news due to its attraction and novelty.[21]

---

[10] Facebook Community Standards Enforcement Report 2022 'Hate Speech' available at *https://transparency.fb.com/data/community-standards-enforcement/hate-speech/facebook/*, accessed on 1 October 2022.
[11] Facebook Community Standards Enforcement Report 2022 'Violent and Graphic content', available at *https://transparency.fb.com/data/community-standards-enforcement/graphic-violence/facebook/*, accessed on 1 October 2022.
[12] Facebook Community Standards Enforcement Report 2022 'Bullying and Harassment', available at*https://transparency.fb.com/data/community-standards-enforcement/bullying-and-harassment/facebook/*, accessed on 1 October 2022
[13] Tarleton Gillespie 'Platforms are not intermediaries' (2018) 2.2 *Georgetown Law Technology Review 1*98 at 207.
[14] Ibid.
[15] Grimmelmann op cit note 9 at 54.
[16] Lior J Strahilevitz 'Wealth without markets?' (2007) 116 *Yale Law Journal* 1472 at 1493.
[17] Grimmelman op cit note 9 at 54.
[18] Maria D Molina, S. Shyam Sundar, Thai Le & Donwon Lee '"Fake News" is not simply false information: A concept explication and taxonomy of online content' (2021) 65(2) *American Behavioural Scientist* 180 at 180.
[19] Alvaro Figueira & Luciana Oliveria 'The current state of fake news: Challenges and opportunities' (2017) 121 *Procedia Computer Science* 817 at 818.
[20] Molina et al op cit note 18 at 183.
[21] Alexandre Bovet & Hernan Makse 'Influence of fake news in twitter during the 2016 US Presidential Election' (2019) 10 *Nature Communications* 1 at 2.

The consequences of these harms are not confined to the online realm and examples abound of their effects in real life. GamerGate is an example of persistent and coordinated harassment targeted at prominent women in the gaming industry. They were subjected to months of online abuse that included threats of rape, physical harm and death.[22] One woman fled her home after her address was posted online and another went into hiding.[23]

Such harm is also not confined to the physical but can permeate and undermine important societal functions, values, and rights. This is evidenced by the alleged impact of false information on the outcome of elections.[24] The orchestrated dissemination of false information at important moments shapes public discourse which may influence the outcome of democratic decisions.

Such harms are further felt by a shrinking marketplace of ideas and the undermining of the right to freedom of expression. The proliferation of harmful content and behaviour contributes to a toxic culture on platforms.[25] Consequently, many users may censor themselves, avoid particular content or leave platforms.[26] This has a chilling effect on freedom of expression and undermines the potential of platforms to flourish as inclusive online communities.[27]

(b) The rationale for moderation and how it is defined

Content moderation is posited as a solution to these myriad harms.[28] Prompting some to argue that it is the difference between platforms that succeed and those that are abandoned.[29]

---

[22] Kate Klonick 'The new governors: the people, rules and processes governing online speech' (2018) 131 *Harvard Law Review* 1598 at 1628.
[23] Emily VanDerWerff '#Gamergate: Here's why everybody in the video game world is fighting' *Vox* 2014, available at *https://www.vox.com/2014/9/6/6111065/gamergate-explained-everybody-fighting*, accessed on 17 February 2021.
[24] See for example reports on the 2016 United States of America election and the 2022 Kenyan election. Following the 2016 USA election, it was reported that Russian operatives allegedly created 50, 258 Twitter accounts which were linked to bots (fake accounts which were programmed to disseminate false information). See Sophie Marineau'Fact check US: what is the impact of russian interference in the US presidential election?' *The Conversation* 29 September 2020 available at https://theconversation.com/fact-check-us-what-is-the-impact-of-russian-interference-in-the-us-presidential-election-146711, accessed on 26 November 2020.
[25] Gillespie op cit note 13 at 207.
[26] Ibid.
[27] Gillespie op cit note 6 at 16.
[28] Kyle Langvardt 'Regulating online content moderation' (2018) 106 *Georgetown Law Journal* 1353 at 1358.
[29] Grimmelmann op cit note 9 at 45.

This is the rhetoric associated with moderation – it is conducted to prevent harm. Grimmelmann accepts this motivation and includes it in his definition of moderation which he considers 'the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse.'[30] However, several other factors influence the moderation policies of platforms including regulatory compliance, profit maximisation and response to public outcry.[31] Platforms profit by collecting large amounts of data about users to sell targeted advertising.[32] They accordingly have an interest in increasing engagement because it creates more data. This motivates platforms to promote content that is 'emotionally charged, extreme and inflammatory'[33] because it spreads quickly and is viewed extensively. The prevention of harm is accordingly not the only rationale for moderation, it may be intricately tied with business models that prioritise user engagement, data surveillance and profit. [34] Because of this, I define moderation more neutrally.

I define it as the governance mechanisms that determine the bounds of permissible content and the actions that may be taken in response to violations. In practical terms, moderation determines what content to remove, what to highlight and what to suppress on a platform. Such decisions dictate the marketplace of ideas, founding perceptions that platforms are 'content gatekeepers[35], 'governors of online speech'[36] and 'custodians of the public sphere'.[37] Such a process is at odds with early ideals of the internet but is now widely conducted by platforms.

(c) The operation of content moderation

The operation of moderation entails the development of rules, the creation and training of systems to enforce the rules and a process of iteration that relies on external reaction to update them.[38] It also involves a process to adjudicate on compliance with such rules, and responses to non-compliance. Such a system is analogous to a governance or legal system[39] and its operation has

[30] Ibid at 47.
[31] Barrie Sander 'Freedom of expression in the age of online platforms: the promise and pitfalls of a human rights-based approach to content moderation' (2020) 43 *Fordham International Law Journal* 939 at 950–955.
[32] Daphne Keller 'Internet platforms: observations on speech, danger, and money' (2018) *Hoover Working Group on National Security, Technology and Law, Aegis Series Paper No. 1807* at 4.
[33] Sander op cit note 31 at 954.
[34] Ibid.
[35] Ibid at 941.
[36] Klonick op cit note 22 at 1602.
[37] Jack M Balkin 'Free Speech is a Triangle' (2018) 118 *Columbia Law Review* 2011 at 2022.
[38] Ibid.
[39] Klonick op cit note 22 at 1602.

aptly been characterised as a form of 'private governance.'[40] This conception is grounded on the 'quasi-normative', 'quasi-executive' and 'quasi-judicial' powers exerted by platforms when they make the rules, enforce the rules, and determine compliance with the rules.[41]

(i) Quasi-normative powers

Platforms provide the rules for moderation in publicly accessible documents referred to as 'community standards' or 'terms of use.'[42] Users are typically required to accept the terms to access the platform.[43]The rules dictate the type of content that is permitted and prohibited. Each platform has different rules which are underpinned by the specific values of the platform's community.[44] Although there are distinctions between platforms, there is significant overlap. For example, Facebook, Twitter and TikTok, all prohibit the sexual exploitation of children and have rules concerning hate speech and violence.

The rules generally apply to all users, regardless of their location as expressly provided by Facebook[45] and TikTok.[46] However, some platforms treat content differently depending on the jurisdiction. Twitter withholds content in certain countries if they receive a request from 'an authorised entity' (such as a court order) or the content violates local laws.[47] Access will only be restricted in that territory. Access may also be age restricted.[48]

[40] Balkin op cit note 37 at 2032.
[41] Luca Belli, Pedro Augusto Francisco & Nicolo Zingales 'Law of the land or law of the platform? Beware of the privatisation of regulation and police' in Luca Belli & Nicolo Zingales (eds) *Platform regulations (2017)* 41at 59.
[42] Facebook's content rules are referred to as Community Standards and TikTok refers to them as Community Guidelines.
[43] Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression Doc A/HRC/32/38 11 May 2011.
[44] Klonick op cite note 22 at 1628.
[45] Facebook's Community Standards, available at *https://transparency.fb.com/en-gb/policies/community-standards/*, accessed on 1 October 2022.
[46] TikTok's Community Guidelines, available at *https://www.tiktok.com/community-guidelines?lang=en*, accessed on 1 October 2022.
[47] Twitter Help Center 'About country withheld content', available at *https://help.twitter.com/en/rules-and-policies/tweet-withheld-by-country*, accessed on 1 October 2022.
[48] Facebook will only make certain content available for users over 18. See for example, Facebook's Community Guidelines on content concerning adult nudity and sexual activity.

Such rules also typically apply to all forms of expression–including videos, livestreams, pictures, audio, links and text-based expression.[49] Twitter enforces its rules against tweets, which are publicly accessible, as well as direct messages, which are private.[50]

Such rules *determine the bounds of permissible content* by dictating what can and cannot be expressed. This forms the first part of our definition. The second part – *the actions that may be taken against infringing content* – includes the measures that platforms take against content that violates their rules.

(ii) Quasi-executive powers

Moderation policies are enforced in several ways including removing the impugned content, restricting access to it, temporarily or permanently suspending a user's account, or temporarily or permanently banning someone from the platform.[51] The action taken by a platform is generally tiered, depending on its severity. This is evident on Facebook where adult nudity and sexual activity content, is either removed, placed behind a sensitivity-warning or restricted for individuals under 18.[52] On Facebook, TikTok and Twitter, suspension or banning is reserved for instances of severe violations.[53] In making such determinations they consider whether the user repeatedly violated the rules, and TikTok considers 'actions on other platforms and offline behaviour.'[54]

Platforms also respond to violating content by reducing its visibility. TikTok's Community Guidelines state 'for some content, we may reduce discoverability, including by redirecting search results, or making videos ineligible for recommendation.'[55] This appears to be a highly discretionary response. This is the case on Twitter too. Twitter provides very little information on when or how decisions are made, and simply notes: '[l]imiting Tweet visibility depends on a number of signals about the nature of the interaction and the type of content.'[56] These provisions

[49] TikTok's Community Guidelines explicitly state that it applies to 'everything on TikTok'.
[50] Twitter op cit note 47.
[51] Facebook op cit note 45.
[52] Ibid.
[53] TikTok for example, provides that they will 'ban accounts and/or users that are involved in severe or repeated on-platform violations.' The banning of Ex-President Donald Trump from Twitter is a notorious example of this.
[54] TikTok op cit note 46.
[55] Ibid. Additional information is provided under 'Ineligible for the For You-Feed' which provides the types of content which would be considered ineligible, these include: dangerous stunts and sports, overly sexualised content and tobacco and alcohol products, amongst others.
[56] Twitter op cit note 47.

allow platforms to choose what information to highlight, and what information to suppress; ultimately determining what expression is valuable, and what shouldn't be seen. Such decisions underpin conceptions that they curate discourse.

The policies are enforced using a combination of human intervention and automation, and there are different processes that may lead to action. [57] Content may be automatically flagged by software and removed before it is ever published. [58] This is a form of prior restraint. It may be published but then flagged as violating content by software or users, which is then reviewed.

(iii) Quasi-judicial powers

The wording of the rules affords platforms a high degree of discretion – to determine what expression amounts to, whether it is a violation, and the appropriate response. Importantly, Twitter, TikTok and Facebook all have a public interest override. This permits violating expression to remain online if it is in the public interest. This is not an objective determination, and all three platforms provide vague substantiation of its application. TikTok includes an exhaustive list of the circumstances that warrant a deviation from removal for offending content, which includes satire, counter-speech and content that 'enables individual expression on topics of social importance.'[59] It is not clear what this means. In determining the public interest, Facebook weighs up whether there is a threat to safety if it is a voice in current debates, circumstances in the country, the type of speech and the country's political structure.[60]

Users are typically notified of the action taken in response to their content. But even this is discretionary. Facebook provides that notice will not be provided if doing so would attract liability, harm their community, or technical limitations restrict them from doing so.[61] Harm to their community could mean any number of things, as could a technical limitation.

[57] Grimmelman op cit note 9 at page 67.
[58] In this regard, see Facebook's Transparency Center 'How technology detects violations', available at *https://transparency.fb.com/en-gb/enforcement/detecting-violations/technology-detects-violations/*, accessed on 2 October 2022.
[59] TikTok op cit note 46.
[60] Facebook op cit note 45.
[61] Facebook's Terms of Service (last revised 26 July 2022), available at *https://www.facebook.com/terms.php?ref=p*, accessed on 1 October 2022.

Importantly, social media platforms generally provide users with a review mechanism. This enables a user to request a review in instances where their content has been incorrectly removed or their account is disabled or deleted.[62] Facebook has even appointed an Oversight Board – an independent body that is mandated to hear select appeals.[63]

It is evident why such powers are akin to the exercise of public power. However, there is an important distinction to highlight – platforms are not constrained in their exercise of power in the same way that public bodies are. Their normative powers do not need to comply with a constitution, which includes a Bill of Rights.[64] Nor does such rule-making prescribe to the standard principles of transparency, accountability and consistency.[65] Their quasi-executive power is not constitutionally required to be lawful, reasonable or procedurally fair.[66] Their quasi-judicial power is not informed by precedent or exercised by adjudicators who are bound by professional ethics and principles of independence and impartiality.[67]

This important consideration sits amongst a myriad more, which are briefly detailed below.

(d) Considerations for content moderation

In this section, we explore four considerations that highlight the difficulties of content moderation and the context within which it operates.

First, moderation is hard. In February 2022, the Russian government ordered Meta to stop fact-checking and labelling content posted by Russian media sites.[68] Meta refused, and Russia announced it was restricting access to the platforms. Meta's platforms were being used by Russian citizens to express and organise themselves during a crisis.[69] In essence, Meta had to decide between allowing a government to disseminate any content it liked – possibly false – or keeping

[62] Ibid.
[63] See Facebook op cit note 58.
[64] As is provided for in section 2 of the Constitution of the Republic of South Africa, 1996.
[65] Emily B Laidlaw 'Online platform responsibility and human rights' in Luca Belli & Nicolo Zingales (eds) *Platform regulations (2017)* 65 at 69.
[66] As is provided for by section 33 of the Constitution.
[67] As is provided for in section 165(4) of the Constitution.
[68] Taylor Hatmaker 'Russia Says it is Restricting Access to Facebook in the Country' *Techcrunch*,25 February 2022 available at *https://techcrunch.com/2022/02/25/russia-facebook-restricted-censorship-ukraine/*, accessed on 3 March 2022.
[69] Ibid.

its platforms accessible. An impossible choice. This demonstrates the difficult role that platforms play in finding the appropriate balance between mitigating harm and enabling freedom of expression. Such difficulties are heightened during a developing geopolitical crisis or global pandemic.[70]

Second, platforms are influenced by multiple stakeholders, with various motives.[71] They do not exist in a vacuum and are subject to significant external pressures. Governments use the threat of restriction or ban to force moderation in line with domestic agendas. This is evidenced by the example above, as well as TikTok's receipt of 4 156 government requests to restrict or remove content in 2021.[72] This is problematic when used improperly by governments to quell opposition or suppress political criticism. It may be difficult for platforms to determine whether such requests are justified or amount to censorship. This also applies to the user take-down process, where abusive requests are made to silence other users.[73]

Third, moderation entails sifting through large quantities of information - 1.3 million posts are shared every minute on Facebook.[74] This has necessitated the use of automated technologies which can sift through high volumes at speed. However, their use has resulted in additional challenges. Their process, which removes content before it is published, creates a mechanism for prior restraint, where technology makes snap decisions.[75] The black-box nature of artificial intelligence makes it difficult to understand why a decision was taken, and consequently how to challenge it. This is underscored by critiques that the decisions are based on discriminatory assumptions and biased datasets.[76]

Fourth, the globalised nature of the internet makes it difficult to shape rules that appropriately account for local contexts.[77] There is no universal definition of hate speech.[78] Moderation policies

[70] Evelyn Douek argues that the Covid-19 pandemic pushed content moderation into a state of emergency. See Evelyn Douek 'Governing online speech: from "posts-as-trumps" to proportionality and probability' (2021) 121 *Columbia Law Review* 759 at 763.
[71] Sander op cit note 32 at 959.
[72] TikTok 'Government Removal Requests Report' 12 May 2022, available at *https://www.tiktok.com/transparency/en/government-removal-requests-2021-2/*, accessed on 1 October 2022.
[73] Keller op cite note 32 at 4.
[74] Laidlaw op cit note 65 at 66.
[75] Langvardt op cit note 28 at 1358.
[76] Sander op cit note 31 at 958.
[77] Balkin op cit note 37 at 2018.
[78] *Qwelane v South African Human Rights Commission* 2021 (6) SA 579 (CC) para 79.

are either too permissive or too restrictive, depending on the history and context of a particular jurisdiction.[79] This also poses jurisdictional challenges for the vindication of rights.

These factors shape the reality that platforms will get it wrong. All moderators – human or artificial – will make mistakes.[80] But it is difficult to know whether a poor decision was a genuine mistake, or whether it was dictated by ulterior motives. The next sections explore the impact that moderation, and its mistakes, have on the right to freedom of expression in South Africa.

III. THE RIGHT TO FREEDOM OF EXPRESSION IN SOUTH AFRICA

(a) Un-packing the right to freedom of expression in South Africa

The right to freedom of expression is provided for in section 16 of the Constitution of South Africa. It holds a significant space in our constitutional democracy because of our history of censorship. That history entailed a regime of broad censorship where political, artistic and cultural expression was restricted.[81] Such a suppressive and intolerant past has garnered a strong appreciation for the importance of the right.[82]

The rationales for the right have been acknowledged by our courts to include its role in the functioning of democracy, in the search for truth and in every person's ability to self-actualise.[83] As noted by the Constitutional Court in *Democratic Alliance v African National Congress:*[84]

> The Constitution recognises that people in our society must be able to hear, form and express opinions freely. For freedom of expression is the cornerstone of democracy. It is valuable both for its intrinsic importance and because it is instrumentally useful. It is useful in protecting democracy, by informing citizens, encouraging debate and enabling folly and misgovernance to be exposed. It also helps the search for truth by both individuals and society generally. If society represses views it considers unacceptable, they may never be exposed as wrong. Open

[79] Sander op cit note 31 at 956.
[80] Douek op cit note 70 at 762.
[81] Pierre De Vos 'Rejecting the free marketplace of ideas: a value-based conception of the limits of free speech' (2017) 33 *SAJHR* 359 at 361.
[82] *Economic Freedom Fighters v Minister of Justice and Correctional Services* 2021 (2) SA 1 (CC) para 2.
[83] *South African National Defence Union v Minister of Defence and Another* 1999 (4) SA 469 para 7.
[84] 2015 (2) SA 232 (CC) paras 122-3.

> debate enhances truth-finding and enables us to scrutinise political argument and deliberate social values.

The Constitutional Court has acknowledged a fourth rationale – 'the encouragement of tolerance.'[85] Tolerance is considered a correlative of freedom of expression. It does not require the acceptance of a particular view, but rather an acceptance that disagreements can be aired in public and that unpopular views will not be silenced.[86] This is underscored by the broad scope of the right which protects unfavourable or controversial expressions too. As noted by the Constitutional Court in *Islamic Unity Convention v Independent Broadcasting Authority:*[87]

> Freedom of expression is applicable, not only to information or ideas that are favourably received or regarded as inoffensive or as a matter of indifference, but also to those that offend, shock or disturb the state or any sector of the population. Such are the demands of that pluralism, tolerance and broadmindedness without which there is no democratic society.

The right to freedom of expression accordingly protects a wide range of expression, including that which may be controversial, or offensive. It is broad in scope, and the Constitutional Court has noted the obligation to interpret the scope of protected expression generously.[88]

(b) The operation and limitations of the right to freedom of expression

Despite its broad scope, the right is not absolute[89] and 'not all speech is created equal.'[90] Our law has placed lawful restrictions on the types of content that enjoy constitutional protection and acknowledges instances where a limitation of the right is justified. This dichotomy is evident in the construction of the right where section 16(1) broadly notes the scope of protection and section 16(2) specifies its limitations.

---

[85] *Qwelane* supra note 78 para 69.
[86] *SANDU* supra note 83 para 8.
[87] 2002 (4) SA 294 (CC) at para 26, which endorsed *Handyside v the United Kingdom*, no 5493/72 § 49, ECHR, 1976.
[88] *Laugh it off Promotions CC v South African Breweries International (Finance) BV* 2006 (1) SA 144 (CC) para 47.
[89] Iain Currie & Johan de Waal *The Bill of Rights Handbook* 5 ed (2005) at 163.
[90] Dario Milo & Pamela Stein *A Practical Guide to Media Law* (2013) at 5.

Section 16(1) explicitly acknowledges certain important elements of the right including freedom of the media,[91] freedom to receive or impart information[92]; artistic creativity[93] and freedom of academic and scientific research.[94]

Section 16(2) lists three categories of expression that are not constitutionally protected. These are propaganda for war;[95] incitement of imminent violence;[96] and advocacy of hatred that is based on race, ethnicity, gender or religion and that constitutes incitement to cause harm.[97] The exclusion of some expression from constitutional protection was justified by the Constitutional Court in *Islamic Unity Convention v Independent Broadcasting Authority* due to its potential to adversely affect the dignity of others and to cause harm.[98]

In effect, this construction means that any regulation of the right which falls outside of section 16(2) and encroaches on the bounds of section 16(1), amounts to a limitation of the right.[99] Such a limitation is only constitutionally permissible if it is considered reasonable and justifiable in terms of section 36(1) of the Constitution.[100]

South Africa's jurisprudence contains many examples of justifiable limitations of the right to freedom of expression. Legislative measures include the Promotion of Equality and Prevention of Unfair Discrimination Act 4 of 2000. It is also lawfully curtailed by the exercise and enjoyment of the rights to dignity and privacy.[101] The reconciliation between such competing rights finds form through the law of delict, specifically the *actio iniuriarum.*[102]

---

[91] Section 16(1)(*a*) of the Constitution.
[92] Section 16(1)(*b*) of the Constitution.
[93] Section 16(1)(*c*) of the Constitution.
[94] Section 16(1)(*d*) of the Constitution.
[95] Section 16(2)(*a*) of the Constitution.
[96] Section 16(2)(*b*) of the Constitution.
[97] Section 16(2)(*c*) of the Constitution.
[98] *Islamic Unity* supra note 87 para 30
[99] Ibid para 34.
[100] Currie op cit note 89 163.
[101] Jonathan Burchell 'The legal protection of privacy in South Africa: A transplantable hybrid' (2009)13 *Electronic Journal of Comparative Law* 1 at 1.
[102] Debbie Collier 'Freedom of expression in cyberspace: Real limits in a virtual domain' (2005) 1 *Stell LR,* 21 at 28.

(c) Freedom of expression online

Platforms are important spaces for the exercise of the right to freedom of expression. Billions of people around the world use them to share expression, and access information. They have accordingly become regarded as the new marketplace for speech and ideas.[103] In the American case of *Packingham v North Carolina,*[104] Justice Kennedy described social media platforms, as the 'modern public square'[105] and noted that denying an individual access to them restricts their ability to exercise their free speech rights. According to him, the internet, and platforms, are the most important spaces for the exchange of ideas.[106]

This is true in South Africa too – there are currently over 41 million internet users.[107] In 2022, there were an estimated 28 million social media users which equates to nearly half the population.[108] A significant amount of our population is accordingly exercising their rights on the internet and social media platforms specifically. Our courts have acknowledged this.

In *Democratic Alliance v African National Congress*[109] the Constitutional Court included the internet as a space for vigorous public debate, alongside public meetings, radio, television and newspapers.[110] In *Afriforum v Pienaar,*[111] the Western Cape High Court extended the freedom of the press and other media to social media platforms[112] and in *Tschilas v Touch Line Media* the court acknowledged the role that chat forums play in enabling debate,[113] and focused on the importance of freedom of expression on the internet.

Human rights apply online;[114] implicating the considerations and limitations discussed. Online expression can amount to hate speech and result in consequences in the same way that it could if published in a newspaper. Our courts have navigated this by developing existing legal mechanisms

[103] Klonick op cit note 22 at 1613.
[104] 137 S. Ct. 1730 (2017).
[105] Ibid.
[106] Ibid at 1735 (quoting *Reno v ACLU*, 521 U.S. 844, 868 (1997).
[107] Datareportal 'Digital 2022: South Africa', available at *https://datareportal.com/reports/digital-2022-south-africa*, accessed on 28 September 2022.
[108] Ibid.
[109] 2015 (2) SA 232 (CC).
[110] Ibid para 134.
[111] 2017 (1) SA 388 (WCC).
[112] Ibid para 55.
[113] *Tsichlas v Touch Line Media (Pty) Ltd* 2004 (2) SA 112 (W) at 119A-C.
[114] Special Rapporteur op cit note 4 at 4.

to account for expression on the internet.[115] This is appropriate for disputes that concern the implications of one speaker's expression on another's rights; and where South African courts ordinarily have jurisdiction.[116] However, the structure of the regulation of expression has changed.[117] It doesn't simply concern the state and the speaker. It concerns additional players including platforms, internet service providers and trolls.[118] Our courts have not adjudicated a matter that concerns the implications of these new players' actions on the freedom of expression of a South African user. So what is their impact? The next section analyses how content moderation undermines the right to freedom of expression.

IV. HOW CONTENT MODERATION UNDERMINES THE RIGHT TO FREEDOM OF EXPRESSION IN SOUTH AFRICA

Content moderation undermines the right to freedom of expression in South Africa. This section discusses how, by exploring four possible grounds of unconstitutionality.

Content moderation policies specify the rules that govern permissible expression on platforms. These rules regulate expression. In South Africa, regulation is not required to give effect to the right to freedom of expression.[119] Because of this, any regulation of the right is considered a limitation.[120] Such regulation is not always fatal– it may be considered reasonable and justifiable in terms of section 36 of the Constitution – and will accordingly stand. However, such an analysis would fall short on the first hurdle every time because a moderation policy is not a law of general application. The purpose of this section is not to determine the justifiability of each limitation but rather to illustrate examples of how content moderation policies may undermine the right.

---

[115] Daniel Sive & Alistair Price 'Regulating expression on social media' (2019) 136 *SALJ* 51 at 67. See for example, *The Chinese Association Gauteng v Henning and Others* (EQ2/2017) [2022] ZAGPHJC 590 (28 July 2022), which concerned expression on Facebook, some of which amounted to hate speech and was found to contravene the Equality Act.

[116] In the matter of *Chinese Association* supra note 115, for example, all of the parties fell within the jurisdiction of the court.

[117] Jack M Balkin conceives of this new structure as a triangle. See Balkin op cit note 37 at 2015.

[118] Ibid.

[119] *Print Media South Africa v Minister of Home Affairs* 2012 (6) SA 443 (CC) para 51.

[120] Ibid para 51. However, it must be noted that regulation of the types of expression provided for in section 16(2) of the Constitution would not be considered a limitation of the right in section 16(1), see *Qwelane* supra note 78 para 76.

They may do so, for four reasons. First, they prohibit categories of protected expression that are not included in section 16(2) of the Constitution. Second, they are vague and fall short of the requirements of the rule of law.[121] Third, the process of moderation entails prior restraint and fourth, their application extends to the private sphere.

(a) The limitations extend beyond section 16(2)

Facebook, TikTok and Twitter all place limitations on categories of expression that would ordinarily be constitutionally protected. Facebook includes limitations on expression concerning adult nudity and sexual activity,[122] and suicide and self-injury.[123] TikTok limits 'disordered eating' and violent and graphic content and Twitter restricts content that suppresses participation in civic processes.[124] All three platforms prohibit or place limitations on false information. [125] [126]

These categories of expression are not listed in section 16(2) of the Constitution and are accordingly considered protected expressions in terms of section 16(1). Their restriction amounts to a limitation of the right.

(b) The limitations are vague

Section 1(c) of the Constitution provides that South Africa is founded on the rule of law. This requires that rules must be sufficiently clear to enable individuals to comply. In this regard, the Constitutional Court has held:

> [The rule of law] requires that laws must be written in a clear and accessible manner. What is required is reasonable certainty and not perfect lucidity. The doctrine of vagueness does not require absolute certainty of laws. The law must indicate with reasonable certainty to those who are bound by it what is required of them so that they may regulate their conduct accordingly.[127]

[121] As is provided for in section 1(c) of the Constitution.
[122] Facebook op cit note 45.
[123] Ibid.
[124] The Twitter Rules, available at *https://help.twitter.com/en/rules-and-policies/twitter-rules*, accessed on 28 September 2022.
[125] Facebook regulates 'misinformation', see Facebook op cit note 45.
[126] TikTok regulates 'harmful misinformation' and 'integrity and authenticity', see TikTik op cit note 46.
[127] *Affordable Medicines Trust v Minister of Health* 2006 (3) SA 247 (CC) para 108.

The Court has further provided that for a rule to be unconstitutionally vague – it must be more than just poorly constructed – it must be completely 'meaningless and unworkable.'[128]TikTok regulates misinformation by providing:

> [m]isinformation is defined as content that is inaccurate or false. We will remove misinformation that causes significant harm to individuals, our community, or the larger public regardless of intent.[129]

This limitation is inherently vague – what is false? Such a classification necessarily entails universal truth and agreement on such truth.[130] Facebook acknowledges the impossibility of this standard by noting that 'what is true one minute may not be true the next minute.'[131] This is of particular concern during developing or unprecedented events where there are no previous facts to verify content[132] or where there are no authoritative sources to defer to.[133] The wording poses the following difficulties - would misleading information (which is not false) be considered inaccurate? Would genuine opinion, which is identified as such, fall within the ambit of this section if those making determinations disagreed with it? What happens when authoritative sources contradict each other?[134]

As evidenced by these concerns, the wording of this limitation is not sufficiently clear to enable a user to know, with certainty, whether their expression falls foul of the provision. Vague and overly broad provisions, of which there are many, limit the right to freedom of expression and may result in users' self-censoring, which has a 'chilling effect' on the right to freedom of expression.[135]

(c) Prior restraint

The operation of content moderation may, in some instances, entail the censorship of content before it is published on a platform. This process amounts to a system of prior restraint.[136] It occurs when content is automatically flagged and never published. Prior restraints have been criticised

[128] *National Credit Regulator v Opperman* 2013 (2) SA 1 (CC) para 46.
[129] TikTok op cit note 46.
[130] Molina op cit note 18 at 183.
[131] Facebook op cit note 45.
[132] Molina op cit note 18 at 183.
[133] Douek op cit note 70 at 762.
[134] Ibid.
[135] *Print Media* supra note 119 para 23
[136] Langvardt op cit note 28 at 1359.

for undermining the right to freedom of expression and making expression beholden to the government or an administrative body.[137] It is contended that such mechanisms cause self-censorship, have a chilling effect on freedom of expression and undermine public discourse.[138]Because of this, our courts have adopted a cautious approach to prior restraint.[139]

In *Print Media South Africa v Minister of Home Affairs*[140]*,* the Constitutional Court held that such a mechanism unjustifiably limited the right to freedom of expression and was unconstitutional. In making such a determination, the court noted that the nature of prior restraint entails a shift of control from the right bearer to an administrative body.[141] It noted:

> By investing an administrative body with the exclusive power to grant permission to publish certain material, as well as the power to punish for denying it the opportunity to do so, what is engendered is a scheme in which expression must be justified before, and as a necessary condition for, its release into the public realm.[142]

Importantly, this type of restraint excludes the possibility of scrutiny and engagement from the public and is likely to lead to over-restriction of content.[143] Publications are more likely to be restricted before publication than after, and bodies that are obligated and incentivised to classify and restrict content are more likely to do so.[144] Our courts accordingly only permit it in very narrow circumstances.[145] Such a scheme used by platforms may undermine the right to freedom of expression and be considered unconstitutional.

(d) Restrictions apply to private expression

[137] George Devenish 'Prior judicial restraint and media freedom in South Africa – Some cause for concern' (2011) 74 *Journal of Contemporary Roman-Dutch Law* 12 at 13.
[138] Ibid.
[139] See *Government of the Republic of South Africa v Sunday Times Newspaper* 1995 (2) SA 221 (T).
[140] Supra note 119 para 51.
[141] *Print Media* supra note 119 para 58.
[142] Ibid.
[143] Ibid para 59.
[144] Ibid.
[145] *Midi Television (Pty) Ltd v Director of Public Prosecutions* 2007 (5) SA 540 (SCA) para 15.

The content moderation rules of some platforms apply to private messages.[146] This is arguably necessary to guard against targeted harassment or hate speech. However, the Constitutional Court has held that even lawful prohibitions on hate speech do not permeate private conversations.[147] Providing, that extending them to the private realm is not consistent with the purpose of such prohibitions which are aimed at limiting public expressions of hate speech.[148] Accordingly, the operation of content moderation policies to private messages may be found to unjustifiably limit the right to freedom of expression.

V. THE ROLE OF THE STATE

(a) Why involve the state?

Social media platforms are owned by private, international companies.[149] They develop the rules of engagement for access to their platforms which are accepted by users, who are also private actors. Such is the legal, contractual arrangement governing access to and use of social media platforms. Private international companies are not enjoined to protect and promote the rights in the South African Bill of Rights.[150] So why does this relationship invoke the role of the state at all?

The answer is that it doesn't have to, but it should. States all over the world are waking up to the fact that content moderation shapes public discourse and impacts human rights.[151] What to censor is an important public policy consideration, and in determining that, platforms exercise powers akin to public or state power.[152] Because of this, it is contended that states have a duty to constrain the exercise of such power[153] and there have been calls for public scrutiny to guard against

[146] Twitter notes that its policies apply to direct messages, which are private. See in this regard Twitter's Help Center 'Our range of enforcement options' available at *https://help.twitter.com/en/rules-and-policies/enforcement-options*, accessed on 2 October 2022.
[147] *Qwelane* supra note 78 para 118.
[148] Ibid.
[149] For example, Meta is a private company, incorporated in the USA, that owns Facebook, Instagram and Whatsapp, amongst others.
[150] Kyle Langvardt remarks that '[t]hat Facebook is not a governmental actor, of course, relieves all formal constitutional concerns about the company's content restriction policies.' See Langvardt op cit note 28 at 1357. See also *Kaunda v President of the Republic of South* Africa 2004 (10) BCLR 1009 (CC) para 37 which held that the Bill of Rights does not have general application outside South Africa.
[151] Ibid.
[152] Ibid.
[153] Belli op cit note 41 at 60.

abuse.[154] Lucca, for example, remarked that 'it is simply unacceptable for States to throw their hands up and let platforms define the content, scope and limitations of fundamental rights without adequate constraints.'[155]

There is accordingly a growing shift away from the view that the private nature of social media companies insulates them from a duty to respect and protect the rights of their users.[156] International proponents contend that such a view is supported by the Human Rights Council's adoption of the United Nations' Guiding Principles on Business and Human Rights[157] which provides a framework for private companies to respect and protect human rights. Further, international human rights law requires states to ensure that human rights violations are not committed by private parties.[158] Specifically, the state's obligation to respect, protect and promote human rights extends to the oversight of private companies.[159] This view is supported by the Council of Europe which recently remarked that placing these considerations squarely within the realm of contract would be 'untenable' considering jurisprudence from the European Court of Human Rights.[160]

What this jurisprudence shows is an acknowledgement that the internet and platforms are vital for the exercise of human rights, specifically freedom of expression.[161] Their governance cannot be viewed only in terms of contract. It also shows a reimagining of the right to freedom of expression to include positive obligations. Such obligations require states to take certain domestic measures to ensure the right may be exercised and is not interfered with by private, international persons.[162] In *Dink v Turkey*[163] the court pronounced that states have a duty to ensure favourable conditions

[154] Ibid.
[155] Ibid.
[156] Evelyn Douek 'The limits of international law in content moderation' (2021) 6 *UC Irvine Journal of International, Transnational and Comparative law* 37 at 40.
[157] Human Rights Council Resolution 17/4 United Nations Document A/HRC/Res/17/4 (6 July 2011); John Ruggie (Special Representative of the Secretary-General), Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework, UN Doc A/HRC/17/31 (21 March 2011).
[158] Aleksandra Kuczerawy 'The power of positive thinking: Intermediary liability and the effective enjoyment of the right to freedom of expression' (2017) *KU Leuven Centre for IT & IP Law CiTiP Working Paper 30/2017* at 6.
[159] United Nation's General Comment No 31 [80] adopted on 29 March 2004, available at *https://www.refworld.org/docid/478b26ae2.html,* accessed on 19 October 2022.
[160] Council of Europe 'Guidance note on content moderation' (2021) adopted by the Steering Committee for Media and Information Society Content Moderation at 21.
[161] See *Cengiz v Turkey,* nos 48226/10 and 14027/11 (2015) para 49.
[162] Kuczerway op cit note 159 at 8.
[163] *Dink v Turkey* (2010) ECtHR.

exist to enable everyone to exercise their right to freedom of expression. In *Özgür Gündem v Turkey,* the European Court of Human Rights held that the exercise of the right 'does not depend merely on the State's duty not to interfere but may require positive measures of protection.'[164] This reimagining enables individuals to hold the state accountable if it has not taken appropriate domestic measures to constrain the powers of private, international companies. Given the impact of content moderation, such inadequate measures would undermine their duty to respect, protect and promote the right to freedom of expression.

(b) Reconceptualising the right in South Africa

Such a reconceptualisation is supported in South Africa's jurisprudence. The right is traditionally considered a negative one; requiring the state to not interfere with it.[165] However, our Courts have acknowledged that even civil and political rights may confer a positive duty.[166] Such argument is premised on section 7(2) of the Constitution which enjoins the state to 'respect, protect, promote and fulfil the rights in the Bill of Rights.'[167] This was recognised by the technical committee working on the Constitution in 1996. In a commentary on section 7 of the Constitution, it noted:[168]

> [t]the obligation 'to protect' the rights requires States to take positive steps to prevent a right from being infringed by both State and private actors. The obligation 'to promote' the rights refers to the duty of the State to take steps to create the necessary culture and social conditions in which the full enjoyment of human rights is possible.

There is accordingly a duty on the state to ensure that rights are not infringed by private actors. Further, states must take necessary measures to ensure that rights may be exercised and enjoyed.[169] This means that the state should use its resources and powers to ensure that the conditions exist for freedom of expression to be exercised. This would include its exercise online. Such positive

[164] *Özgür Gündem v Turkey (*2000) ECtHR para 43.
[165] Currie op cit note 89 at 571.
[166] See *August v Electoral Commission 1999 (3) SA 1 (CC)*, which considers the positive obligations enjoined by the right to vote.
[167] See *Minister of Health v Treatment Action Campaign* 2002 (5) SA 721 (CC) at para 39
[168] Memorandum of the Panel of Constitutional Experts and Technical Committee 4, 8 March 1997, available at *https://www.justice.gov.za/legislation/constitution/history/LEGAL/CP008036.PDF*, accessed on 22 October 2022.
[169] De Vos op cit note 81 at 378.

measures could include the adoption of appropriate administrative or legislative measures to achieve this.[170]

Even a strictly contractual or private nature of platforms may not frustrate such an approach. Our courts have demonstrated a willingness to intervene in private, contractual matters in pursuit of principles of reasonableness, fairness and justice. This is evident in B*arkhuizen v Napier*[171] where the court held that freedom of contract may be constrained if its enforcement would undermine such principles. The Constitutional Court noted that to hold otherwise – that a court could never intervene – would tie the hands of justice, which could never be the case.[172] Public policy is now dictated by the Constitution[173] and the Supreme Court of Appeal has acknowledged that a contractual term that limits a constitutional value will not be enforced.[174] Our courts have also intervened in procedurally unfair private, disciplinary matters.[175] It has done so for churches[176] and professional associations.[177] Their intervention has been rationalised on the importance of the consequences – it may cause reputational damage or economic hardship.[178]

The global shift in how moderation and the right are conceptualised is mirrored by shifts in approaches to regulation. Content moderation used to be predominantly governed by self-regulation, with social media companies voluntarily determining and enforcing their own governance structures, through their moderation policies.[179] However, states have become increasingly more involved in the governance of content moderation.[180] They have begun to take positive steps to constrain the power of platforms to protect and promote the right to freedom of expression. States have taken different approaches to regulation and South Africa's approach is detailed below.

---

[170] Ibid.
[171] *Barkhuizen v Napier* 2007 (5) SA 323 (CC) para 73.
[172] Ibid.
[173] Sive op cit note 115 at 75.
[174] *Bredenkamp v Standard Bank of SA Ltd* 2010 (4) SA 468 (SCA) para 50.
[175] Cora Hoexter *Administrative Law in South Africa* 2 ed (2012) at 444.
[176] *Ndara v Umtata Presbytery, Nederduitse Gereformeerde Kerk in Afrika (Transkei)* 1990 (4) SA 22 (Tk).
[177] *Jockey Club v Forbes* 1993 (1) SA 649 (A).
[178] Hoexter op cit note 176 at 444.
[179] Christopher Marsden *Internet Co-regulation: European Law, Regulatory Governance and Legitimacy in Cyberspace* 1 ed (2011) at 212.
[180] Ibid.

VI. SOUTH AFRICA'S REGULATION OF ONLINE EXPRESSION

This section details the regulatory responses that South Africa has taken to freedom of expression online. Such regulation is provided for in the Electronic Communications and Transactions Act 25 of 2002 ('ECTA') and the Films and Publications Act 65 of 1996 ('FPA'). The FPA has recently been amended[181] to bring online content into its remit.[182] The amendments include several vague provisions which pose challenges for its implementation and have drawn significant criticism.[183]

Together, these laws do four things: first, provide conditional limitation of civil liability for platforms. Second, establish a parallel take-down process that enables the removal of content online. Third, establish a complaints mechanism that may result in the removal of content and fourth, impose a classification scheme for online content that amounts to prior restraint.

(a) Conditional limitation of liability

Provision is made to exclude platforms from civil liability for content on their platform. They would accordingly not be held liable for defamation, for example, by hosting or publishing defamatory content. Such a regulatory position acknowledges that platforms have limited control over the content users publish.[184] It is considered necessary to incentivise platforms to retain their service offerings which enables the exercise of freedom of expression.[185]

Such limitation of liability is provided for in chapter XI of ECTA. However, liability is not automatically excluded. It only applies if certain conditions are met. First, the service provider must be a member of a recognised representative body.[186] To be recognised, members must be subject to a code of conduct that includes adequate standards,[187] there must be requirements for membership and the representative body must be capable of enforcing compliance.[188] Second, it

[181] It was amended by the Films and Publications Amendment Act no 19 of 2020, which came into effect on 1 March 2022.
[182] See the definition of 'publication' in section 1 which explicitly includes expressions published using the internet.
[183] See, for example, Noxolo Majavu & Palesa Dlamini 'Films and Publications Amendment Act leaves online content producers hot under the collar' *News24* 5 March 2022, available at *https://www.news24.com/citypress/news/films-and-publications-amendment-act-leaves-producers-hot-under-the-collar-20220305*, accessed on 10 October 2022.
[184] Sive op cit note 115 at 70.
[185] Keller op cit note 32 at 9.
[186] Section 72(*a*) of ECTA.
[187] Section 71(2)(*a*) and (*c*) of ECTA.
[188] Section 71(2)(*d*) of ECTA.

must have adopted and implemented its official code of conduct.[189] To date, no social media platforms have met these conditions.[190] It is not clear whether social media platforms would collectively associate, or agree on a code of conduct.

Additional requirements for immunity apply depending on the operations of the service provider. Social media platforms typically store data, such as a tweet or post, which is developed and shared by a recipient or user of their service. Platforms would accordingly be considered a 'host' in terms of ECTA.[191] For liability to be excluded, hosts must meet three additional requirements. They must not have actual knowledge that the content infringes the rights of someone else[192] or knowledge of the facts or circumstances of the rights infringement,[193] and they must respond to a take-down notice 'expeditiously.'[194]

(b) The take-down notification process

The take-down notification process is provided for in section 77 of ECTA. Any person is competent to file a notification 'of unlawful activity' with the platform. This legislatively enables private individuals to interfere with another person's expression.[195] Service providers are required to expeditiously remove or disable access to the data following receipt of the take-down notification.[196] A platform will not be held liable if it incorrectly removes content in response to a take-down notification.[197]

The limitation of liability scheme and the take-down process provided for in ECTA are problematic for two reasons – they incentivise removal and outsource important decisions about the exercise of rights.

[189] Section 72(*b*) of ECTA.
[190] Department of Communications GN 588 GG32252 of 22 May 2009.
[191] Section 75(1) of ECTA considers a host to be a service provider that 'provides a service that consists of the storage of data provided by a recipient of the service.' Price and Sive concur in the assertion that a social media platform would be considered a 'host' for the purposes of ECTA, see Sive op cit note 115 at 71.
[192] Section 75(1)(*a*) of ECTA.
[193] Section 75(1)(*b*) of ECTA.
[194] Section 75(1)(*c*) of ECTA.
[195] Kuczerawy op cit note 159 at 6.
[196] Section 75(1)(*c*) of ECTA.
[197] Section 77(3) of ECTA.

(i) Incentivising removal

ECTA incentivises platforms to remove content by making the exemption of liability conditional upon an expeditious response to a take-down notification.[198] To avoid liability, a platform will likely act cautiously and remove the content, instead of weighing up any conflicting rights.[199] Such an exercise is burdensome and time-consuming,[200] and they are incentivised by ECTA to respond quickly. If they are slow or do not respond at all, the limitation of liability may not apply to them.[201] To avoid legal and financial risk, a platform is likely to just remove the impugned content, regardless of its legality.[202] Further, if a domestic jurisdiction has legally mandated a notice and takedown system it 'stacks the deck in favour of the accusers.'[203] Platforms are likely to take the request at face value, instead of determining whether the request is legitimate.[204] This is bolstered by the fact that a platform will not be held liable if it incorrectly removes content in response to a take-down notification.[205] In real terms, this means that when confronted with difficult or 'grey-area' expressions such as satire, humour or a developing controversial debate, platforms will simply remove it.[206] Such an approach does not interpret protected expression broadly, or tolerate controversial or offensive expression, as required by our Constitution.[207] This may very well lead to over-removal and the censorship of lawful expression online which undermines the right to freedom of expression.[208]

(ii) Outsourcing rights-based determinations

Such a take-down notification scheme is considered an 'inappropriate transfer of juridical authority to the private sector.'[209] ECTA does not expressly provide which content may be lawfully

[198] Kuczerawy op cit note 159 at 4.
[199] Ibid at 5.
[200] Belli op cit note 41 at 52.
[201] Section 75(1)(c).
[202] Kuczerawy op cit note 159 at 5.
[203] Keller op cit note 32 at 5.
[204] Ibid.
[205] Section 77(3) of ECTA.
[206] Keller op cit note 32 at 5.
[207] *Laugh it off* supra note 88 para 47.
[208] Sive op cit note 115 at 71.
[209] European Commission, Summary of the results of the Public Consultation on the future of electronic commerce in the Internal Market and the Implementation of the Directive on electronic commerce (2000/31/EC), available at *http://ec.europa.eu/internal_market/consultations/docs/2010/ecommerce/summary_report_en.pdf*, accessed on 18 October 2022.

removed, it simply requires that the complainant identify the right that has been allegedly infringed and the content that is the subject of unlawful activity.[210] Defamatory content may infringe on the right to dignity, but it may also be lawful if it is true and in the public interest.[211] Rights infringements that occur because of expression generally entail consideration of the two opposing rights at play. It is not enough to simply allege the existence of a rights infringement to justify the removal of content. As noted above, ECTA doesn't incentivise careful weighing up.

ECTA does not require a platform to hear from the content creator or adhere to the principle of *audi alteram partem* in any other way. Nor does it place any transparency or accountability requirements on the platform to enable a state body to oversee the implementation and outcomes of the process. Further, ECTA does not establish a review mechanism, and there are accordingly no due process requirements included in the law, despite its transfer of juridical authority. Following a platform's internal review process will likely result in the platform simply advising that they were complying with domestic law.

(c) The complaints mechanism

Section 18E of the FPA establishes a complaints mechanism that allows any person to lodge a complaint with the Film and Publication Board ('Board') about online content that is unclassified, prohibited or potentially prohibited.[212] The content that is prohibited includes propaganda for war;[213] content that incites imminent violence;[214] or 'advocates hatred based on any identifiable group characteristic and that constitutes incitement to cause harm and imminent violence.'[215] These prohibitions largely mirror the categories of unprotected expression listed in section 16(2) of the Constitution.[216] It also includes child pornography.[217] If the Board determines the content is unclassified or constitutes prohibited content, it may issue a take-down notice to a platform as

[210] Sections 77(1)(*c*) and (*d*) of ECTA.
[211] See *Times Media Ltd and Others v Nisselow* [2005] 1 All SA 567 (SCA).
[212] Section 18E(1) of the FPA.
[213] Section 16(2)(*b*) of the FPA.
[214] Section 16(2)(*c*) of the FPA.
[215] Section 16(2)(*d*) of the FPA.
[216] However, notably, the prohibition of the advocacy of hatred extends beyond the scope provided for in section 16(2)(c) of the Constitution by extending the listed grounds and requiring the incitement of harm *and imminent violence*.
[217] Section 1 definition of 'prohibited content' read with section 16(4) and section 18(3) of the FPA.

provided for in ECTA.[218] As detailed above, the platform is likely to respond by removing the impugned expression, without considering the legality of the request.

This complaints mechanism expressly applies to online content[219] and poses two concerns. First, it outsources important juridical decisions to an administrative body. Effectively, the board is empowered to determine whether expression advocates hatred or incites imminent violence. As detailed above, these are difficult decisions and our courts have taken issue with administrative bodies making such decisions which have huge implications for human rights.[220]

Second, a huge amount of content is posted online. It is unclear whether the Board has sufficient resources and capacity to deal with such complaints effectively. This is important to consider in light of the need to remove some content quickly to resolve continuing harm.

(d) The obligation to classify

Section 16(2) of the FPA requires that publications, which include 'any content made available using the internet,'[221] be submitted to the Board for classification before they are distributed. To distribute explicitly includes 'to stream content through the internet, social media or other electronic mediums.'[222] Content published online accordingly falls within the scope of this obligation.

In a media briefing, it was noted that 'non-commercial online distributors' are not required to submit their content for classification before posting it online.[223] This would exclude personal users of social media platforms. However, it is not explicitly provided for in the FPA or the regulations and it is not immediately clear whether someone who posts on social media for private, income generation purposes, such as an 'influencer' would be considered a 'commercial online distributor' or not.

[218] Section 18E(2) of the FPA.
[219] Section 18E of ECTA provides that any person may complain about such content 'in relation to services being offered online by any person.'
[220] See *Print Media* supra note 119.
[221] Definition of 'publication' in section 1 of the FPA.
[222] Definition of 'distribute' in section 1 of the FPA.
[223] Film and Publication Amendment Act media briefing, streamed live on 3 March 2022, available at *https://www.youtube.com/watch?v=SSpSEqw5HYw*, accessed on 20 October 2022.

The consequence of classification is that the content will either be classified as a refused classification,[224] will be restricted[225] or ascribed an age restriction.[226] Content classified as 'refused classification' and 'XX' cannot be broadcast or distributed, and it is an offence to do so.[227]

In effect, these sections create a mechanism for administrative prior restraint, a process that enables the restriction of content before it sees the light of day.[228] As discussed above, our courts are hesitant about such mechanisms noting that 'the free flow of constitutionally protected expression is the rule and administrative prior classification should be the exception.'[229] These provisions may accordingly be subject to constitutional challenge.

Regardless, the uncertainty about the application of this section to ordinary non-commercial users of social media may have a chilling effect on users, undermining their exercise of freedom of expression online.

South Africa's regulatory approach to online expression raises several concerns – it incentivises over-removal, does not require adherence to any due-process principles, introduces a mechanism for administrative prior restraint, poses practical challenges and may chill expression. I submit that such an approach does not extinguish the state's constitutional obligation to respect, protect and promote the right to freedom of expression.[230] In fact, in several ways, it undermines it.

VII. CONCLUSION

The exercise of freedom of expression has been fundamentally changed by the internet. This has reshaped the structure of its regulation – requiring consideration of the role of new, private bodies. A role which has enabled expression, but also undermined it. The consequences of content moderation are enormous and its impact on human rights is significant. This poses a regulatory challenge for states all over the world. Some have risen to the challenge by reimagining freedom of expression to place positive duties on the state to intervene – requiring the enactment of

[224] Section 16(4)(*a*) of the FPA.
[225] Section 16(4)(*b*) of the FPA.
[226] Section 16(4)(*d*) of the FPA.
[227] Section 24A(2) of the FPA.
[228] Devenish op cit note 138 at 13.
[229] *Print Media* supra note 119 para 52.
[230] Section 7(2) of the Constitution.

measures to protect it online. Our current legislative measures do not give affect to the right to freedom of expression. They cannot be said to do so when they incentivise removal and have a chilling effect on expression. Our jurisprudence, however, includes approaches which may be developed to shift our conception of the right to freedom of expression. We may yet rise to the challenge.

Bibliography

Cases

1. *Affordable Medicines Trust v Minister of Health* 2006 (3) SA 247 (CC).
2. *Afriforum NPC & Another v Pienaar* 2017 (1) SA 388 (WCC).
3. *August v Electoral Commission 1999 (3) SA 1 (CC).*
4. *Barkhuizen v Napier* 2007 (5) SA 323 (CC).
5. *Bredenkamp v Standard Bank of SA Ltd* 2010 (4) SA 468 (SCA).
6. *Cengiz v Turkey,* nos. 48226/10 and 14027/11 (2015).
7. *Democratic Alliance v African National Congress* 2015 (2) SA 232 (CC) paras 122-3.
8. *Dink v Turkey* (2010) ECtHR.
9. *Economic Freedom Fighters v Minister of Justice and Correctional Services* 2021 (2) SA 1 (CC).
10. *Government of the Republic of South Africa v Sunday Times Newspaper* 1995 (2) SA 221 (T).
11. *Islamic Unity Convention v Independent Broadcasting Authority and Others* 2002 (4) SA 294.
12. *Jockey Club v Forbes* 1993 (1) SA 649 (A).
13. *Laugh It Off Promotions VV v SAB International (Finance) BV t/a Sabmark International* 2005 (8) BCLR 743 (CC).
14. *Midi Television (Pty) Ltd v Director of Public Prosecutions* 2007 (5) SA 540 (SCA).
15. *Minister of Health v Treatment Action Campaign* 2002 (5) SA 721 (CC).
16. *National Credit Regulator v Opperman* 2013 (2) SA 1 (CC).
17. *Ndara v Umtata Presbytery, Nederduitse Gereformeerde Kerk in Afrika (Transkei)* 1990 (4) SA 22 (Tk).
18. *Packingham v North Carolina* 137 S. Ct. 1730 (2017).
19. *Print Media South Africa v Minister of Home Affairs* 2012 (6) SA 443 (CC).
20. *Qwelane v South African Human Rights Commission* 2021 (6) SA 579 (CC).
21. *Özgür Gündem v Turkey (*2000) ECtHR.
22. *Reno v ACLU*, 521 U.S. 844, 868 (1997).

23. *South African National Defence Union v Minister of Defence and Another* 1999 (4) SA 469.
24. *The Chinese Association Gauteng v Henning and Others* (EQ2/2017) [2022] ZAGPHJC 590 (28 July 2022).
25. *Tsichlas v Touch Line Media (Pty) Ltd* 2004 (2) SA 112 (W) at 119A-C.

Books

26. Christopher Marsden *Internet Co-regulation: European Law, Regulatory Governance and Legitimacy in Cyberspace* 1 ed (2011).
27. Cora Hoexter *Administrative Law in South Africa* 2 ed (2012).
28. Dario Milo & Pamela Stein *A Practical Guide to Media Law* (2013).
29. Emily B Laidlaw 'Online platform responsibility and human rights' in Luca Belli &and Nicolo Zingales (eds) *Platform regulations (2017)* 65.
30. Iain Currie & Johan de Waal *The Bill of Rights Handbook* 5 ed (2005).
31. Tarleton Gillespie *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media* (2018)
32. Luca Belli, Pedro Augusto Francisco & Nicolo Zingales 'Law of the land or law of the platform? Beware of the privatisation of regulation and police' in Luca Belli & Nicolo Zingales (eds) *Platform regulations (2017)* 41.

Journals

33. Aleksandra Kuczerawy 'The power of positive thinking: Intermediary liability and the effective enjoyment of the right to freedom of expression' (2017) *KU Leuven Centre for IT & IP Law CiTiP Working Paper* 30/2017.
34. Alexandre Bovet & Hernan Makse 'Influence of fake news in Twitter during the 2016 US presidential election' (2019) 10 *Nature Communications* 1.
35. Alvaro Figueira & Luciana Oliveria 'The current state of fake news: challenges and opportunities' (2017) 121 *Procedia Computer Science* 180.

36. Barrie Sander 'Freedom of expression in the age of online platforms: the promise and pitfalls of a human rights-based approach to content moderation' (2020) 43 *Fordham International Law Journal* 939
37. Daniel Sive & Alistair Price 'Regulating expression on social media' (2019) 136 *SALJ* 51.
38. Daphne Keller 'Internet platforms: observations on speech, danger, and money' (2018) *Hoover Working Group on National Security, Technology and Law, Aegis Series Paper No. 1807*
39. Debbie Collier 'Freedom of expression in cyberspace: real limits in a virtual domain' (2005) 1 *Stell LR* 21.
40. Evelyn Douek 'Governing online speech: from "posts-as-trumps" to proportionality and probability' (2021) 121 *Columbia Law Review* 759.
41. Evelyn Douek 'The limits of international law in content moderation' (2021) 6 *UC Irvine Journal of International, Transnational and Comparative law* 37.
42. Fawzia Cassim 'Regulating hate speech and freedom of expression on the internet: promoting tolerance and diversity' (2015) 28(3) *South African Journal of Criminal Justice* 303.
43. George Devenish 'Prior judicial restraint and media freedom in South Africa – some cause for concern' (2011) 74 *Journal of Contemporary Roman-Dutch Law* 12.
44. Jack M. Balkin 'Free Speech is a Triangle' (2018) 118 *Columbia Law Review* 2011.
45. James Grimmelmann 'The virtues of moderation' (2015) 17 *Yale Journal of Law and Technology* 42.
46. Jonathan Burchell 'The legal protection of privacy in South Africa: a transplantable hybrid' (2009) 13 *Electronic Journal of Comparative Law,* 1.
47. Kate Klonick 'The new governors: the people, rules and processes governing online speech' (2018) 131 *Harvard Law Review* 1598.
48. Kyle Langvardt 'Regulating online content moderation' (2018) 106 *Georgetown Law Journal* 1353.
49. Lior Strahilevitz 'Wealth without markets?' (2007) 116 *Yale Law Journal* 1472.
50. Maria D Molina, S. Shyam Sundar, Thai Le & Donwon Lee '"Fake News" is not simply false information: a concept explication and taxonomy of online content' (2021) 65(2) *American Behavioural Scientist* 180.

51. Pierre De Vos 'Rejecting the free marketplace of ideas: a value-based conception of the limits of free speech' (2017) 33 *SAJHR* 359.
52. Tarleton Gillespie, 'Platforms are not intermediaries' (2018) 2.2 *Georgetown Law Technology Review* 198.

Reports

53. Council of Europe 'Guidance note on content moderation' (2021) adopted by the Steering Committee for Media and Information Society content moderation.
54. European Commission, Summary of the results of the Public Consultation on the future of electronic commerce in the Internal Market and the Implementation of the Directive on electronic commerce (2000/31/EC), available at *http://ec.europa.eu/internal_market/consultations/docs/2010/ecommerce/summary_report_en.pdf,* accessed on 18 October 2022.
55. Human Rights Council Resolution 17/4 United Nations Document A/HRC/Res/17/4 (6 July 2011); John Ruggie (Special Representative of the Secretary-General), Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework, UN Doc A/HRC/17/31 (21 March 2011).
56. Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression Doc 'A/HRC/17/27' 16 May 2011.
57. Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression Doc 'A/HRC/32/38' 11 May 2011.
58. United Nation's General Comment No 31 [80] adopted on 29 March 2004, available at *https://www.refworld.org/docid/478b26ae2.html,* accessed on 19 October 2022.
59. Memorandum of the Panel of Constitutional Experts and Technical Committee 4, 8 March 1997, available at *https://www.justice.gov.za/legislation/constitution/history/LEGAL/CP008036.PDF*, accessed on 22 October 2022.

Newspaper articles

60. David Gilbert 'Facebook and Instagram are censoring protests against police violence in Nigeria' *Vice* 21 October 2021, available at *https://www.vice.com/en/article/jgqeyg/facebook-is-censoring-protests-against-police-violence-in-nigeria,* accessed on 4 February 2022.
61. David Wicock 'Instagram claims monkey emojis "DON'T breach race rules" as Twitter removes 1,000 racist posts after appalling abuse of England stars' *Daily Mail* 12 July 2021, available at *https://www.dailymail.co.uk/news/article-9780627/Racist-tweeters-knock-door-police-face-force-law.html*, accessed on 4 February 2022.
62. Emily VanDerWerff '#Gamergate: here's why everybody in the video game world is fighting' *Vox* 2014, available at: https://www.vox.com/2014/9/6/6111065/gamergate-explained-everybody-fighting, accessed on 17 February 2021.
63. Noxolo Majavu & Palesa Dlamini 'Films and Publications Amendment Act leaves online content producers hot under the collar' *News24* 5 March 2022, available at *https://www.news24.com/citypress/news/films-and-publications-amendment-act-leaves-producers-hot-under-the-collar-20220305*, accessed on 10 October 2022.
64. Sophie Marineau*,* 'Fact check US: what is the impact of Russian interference in the US presidential election?' *The Conversation* 29 September 2020, available at: https://theconversation.com/fact-check-us-what-is-the-impact-of-russian-interference-in-the-us-presidential-election-146711, accessed on 26 November 2020.
65. Taylor Hatmaker 'Russia Says it is restricting access to Facebook in the country' *Techcrunch* 25 February 2022 available at https://techcrunch.com/2022/02/25/russia-facebook-restricted-censorship-ukraine/, accessed on 26 June 2022.
66. Victoria Elms & Kieran Devine 'Euro 2020: Why is it so difficult to track down racist trolls and remove hateful messages on social media?' *Sky News* 21 July 2021, available at https://news.sky.com/story/euro-2020-why-is-it-so-difficult-to-track-down-racist-trolls-and-remove-hateful-messages-on-social-media-12358392, accessed on 4 February 2022.