The Dynamics of Racism, Antisemitism and Xenophobia on Social Media in South Africa

Thierry Rousset

Gavaza Maluleke

Adam Mendelsohn

University of Cape Town · iYunivesithi yaseKapa · Universiteit van Kaapstad

Kaplan Centre
JEWISH STUDIES

JOHANNESBURG
HOLOCAUST
& GENOCIDE
CENTRE

KONRAD
ADENAUER
STIFTUNG

The Dynamics of Racism, Antisemitism and Xenophobia on Social Media in South Africa

The Dynamics of Racism, Antisemitism and Xenophobia on Social Media in South Africa

A report by
Thierry Rousset,
Gavaza Maluleke and
Adam Mendelsohn

University of Cape Town · iYunivesithi yaseKapa · Universiteit van Kaapstad

Kaplan Centre
JEWISH STUDIES

JOHANNESBURG
HOLOCAUST
& GENOCIDE
CENTRE

KONRAD
ADENAUER
STIFTUNG

Published by Konrad-Adenauer-Stiftung Media Programme Sub-Sahara Africa

60 Hume Road | PO Box 55012
Dunkeld 2196 | Northlands
Johannesburg | 2116
Republic of South Africa

Telephone: + 27 (0)11 214-2900
Telefax: +27 (0)11 214-2913/4

www.kas.de/mediaafrica
Twitter: @KASMedia
Facebook: @KASMediaAfrica

ISBN: 978-0-6397-2550-5 (print)
978-0-6397-3015-8 (e-book)

Download an electronic copy of *The Dynamics of Racism, Antisemitism and Xenophobia on Social Media in South Africa* from www.kas.de/dynamics-of-racism-antisemitism-and-xenophobia

Editing: Bruce Conradie

Design, layout and production: Heath White, ihwhiteDesign
heath@ihwhitedesign.com

Printing: Typo Printing Investments,
Johannesburg, South Africa

Thierry Rousset is a lecturer in the Department of Historical Studies at the University of Cape Town focusing on historical methodologies and Southern African histories. His previous research has ranged from analysing metropolitan representations of settler colonialism, nativisation and racialisation in the context of Tristan da Cunha; analyses of the elephant seal oil trade between Cape Town and the Crozet Islands in the nineteenth century; the use of commando violence in Cape frontier zones; and the use of destitute children as a labour source in the Cape Colony during the nineteenth century.

Gavaza Maluleke is a lecturer in the Department of Political Studies at the University of Cape Town. Her most recent postdoctoral work was in the Democracy, Governance and Service Delivery Unit at the Human Sciences Research Council. She also worked as a postdoc in the Becoming Men Research team at the University of Amsterdam and as a consultant at the United Nations University Institute on Globalization, Culture and Mobility. Her research interests are in digital activism, transnational feminisms, migration, gendered violence, masculinities and media studies in Africa. Her current work focuses on digital activism and gendered violence in Post-Apartheid South Africa.

Adam D. Mendelsohn is Director of the Kaplan Centre for Jewish Studies and Associate Professor of History at the University of Cape Town. The Centre, the only of its kind in Africa, conducts research focused on Jews in southern Africa, past and present. He is the author of *Jewish Soldiers in the Civil War: The Union Army* (2022) and *The Rag Race: How Jews Sewed Their Way to Success in America and the British Empire* (2014). He has co-curated exhibitions at the New-York Historical Society, Princeton University Museum of Art, and the Center for Jewish History.

Foreword

Almost three decades after the end of Apartheid, racism, antisemitism and xenophobia are alive in South Africa. In most instances, these forms of 'othering' are expressed in insidious but subtle ways. On social media, by contrast, racism, antisemitism and xenophobia are often explicit, crude and violent. Very rarely are there any legal consequences for those who engage in hate-mongering.

For some years, we, the directors of KAS Media Africa -- the Media Programme of the German Konrad Adenauer Foundation -- and the Johannesburg Holocaust & Genocide Centre have been troubled by the proliferation of hate on social media in South Africa. We are particularly worried by its implications and consequences for broader society. We need open debate and dialogue in order to be able to address the challenges we collectively face, but so much discussion on social media seems to do harm to the body politic.

Together with our partners at the University of Cape Town and the Kaplan Centre for Jewish Studies, we initiated a study that examines racism, antisemitism and xenophobia on Facebook and Twitter. We were looking for patterns in the mass of posts, tweets, and images that circulate on social media. To do this, we focused on a series of case studies. These demonstrate that political discussion on social media is often dominated by angry and loud voices who too often seem to press the 'send button' before thinking of the consequences, or, in some cases, deliberately aim to sow discord and division.

This report has been prepared for thought leaders, heads of religious institutions, the media, educators and community leaders so that we can collectively alter the damaging dynamics revealed by this study.

We thank all those who contributed to our research, and especially the annotators who went through the daunting and emotionally stressful task of coding posts and messages. May their work encourage reflection on how we can and must ensure that debate and discussion in South Africa is generative rather than destructive.

Christoph Plate
KAS Media Africa

Tali Nates
Johannesburg Holocaust &
Genocide Centre

Table of Contents

Executive Summary

Much political discussion on social media is driven and dominated by the hard right and the hard left. Societal fissures along national, racial and economic lines are exploited by groups such as Afriforum, Operation Dudula, and the Economic Freedom Fighters (EFF), as well as white nationalist and alt-right groups, to expand their social media footprint and amplify their message.

These and other users often seek to create a crude 'us'/'them' binary. Those portrayed as 'others' are stigmatised. Those social media users perceived to not tow the appropriate line are routinely described as sell-outs and race traitors, and positioned as having no standing to have their opinions heard.

A variety of racist tropes recur frequently: the animalisation of members of different racial and national groups, the use of slurs, and claims of imminent threat. For #OperationDudula, the latter typically takes the form of narratives of swamping and criminality by African immigrants. For the white right, it often takes the form of the mobilisation of the trope of 'white genocide' and claims of government complicity in the imagined mass murder of whites. For the EFF, it is the pernicious power of white monopoly capital. These grand narratives allow those propagating these tropes to paint themselves as victims rather than aggressors.

The flashpoints investigated in this study demonstrate how new online communities emerge. Rather than galvanising healthy debate through exposure to different viewpoints, in these case studies social media instead fostered small, polarised communities of like-minded individuals. The intemperate and alienating nature of online political discussion hardened boundaries. Those attempting to cross these lines were routinely attacked in hateful terms.

Paradoxically, the hateful content this produced typically spawned more content as a chorus of supporters and detractors chimed in. By generating attention, the production and publication of hate speech thus, in turn, likely increased the amount of time users spent on online platforms. Given that user engagement is the desideratum of social media companies, and that content moderation is complex

and potentially costly, platforms have been slow to provide adequate and timely moderation, particularly that which is sensitive to local languages and cultures.

Users of social media platforms, moreover, demonstrate considerable savvy in understanding and manipulating to their advantage the dynamics of social media platforms and the inadequacies of content moderation. This includes the deliberate resort to derogatory terminology drawn from local languages so as to evade moderation, provocations designed to draw attention and traffic, the use of fake accounts, and other techniques.

Although some episodes relating to racism can galvanise global responses (as was the case with the killing of George Floyd), much racist content draws on local context, language, and events, and thus does not lend itself to automatised moderation or to commercial content moderators unfamiliar with specific local contexts.

When moderation results in content being removed, the content moderation process is often opaque to users. As it stands, the current moderation process prioritises punishing and policing bad behaviour, instead of encouraging user education.

Though current content moderation processes have been criticised, there are instances when it has proven more effective, as shown in our research on the Israel-Gaza conflict in 2021. Although some users spread antisemitic content, much of the problematic material had been removed by the time we collected our data.

In this instance, the content moderation process may have been eased by the global nature of antisemitism; tropes and stereotypes were quickly recognised by users and commercial content moderators. Further, the recurrent nature of tensions in Israel/Palestine also means the social media platforms have large datasets to train algorithms with, thereby aiding the automatisation of moderation.

This does, however, suggest that when sufficient resources are dedicated to a particular region or issue, inroads can be made in reducing the amount of hate speech on social media. Increasing the number of content moderators, broadening their linguistic reach, ensuring they are embedded within the societies on which their work focuses, and insourcing their labour would go some way towards facilitating this.

Social media platforms will have additional incentive to scale up and professionalise content moderation if they were made to regard the current dynamic as harmful to their commercial interests. For example, public pressure on the advertisers from which social media platforms depend for revenue may incentivise social media platforms to improve content moderation practices. Similarly, regulatory interventions can ensure that platforms are obliged to assume some responsibility for content and enforce basic standards of use.

Given that those of school-going age make up an increasing share of social media users in South Africa, schools should be encouraged to provide guidance to students on appropriate online behaviours. This can be achieved by highlighting the potential harms of posting or sharing problematic content, as well as the mechanisms for flagging problematic content.

The dominance of extreme voices in online political discussion indicates the urgency of addressing some of these problematic dynamics. The demonstrated ability of #OperationDudula to translate online activism into real world action is but one example of the seepage of malign influences from the virtual world into public life. Urgent action is needed.

Policy Recommendations

This report shows how existing societal fissures along racial, national, and economic lines are exploited on social media to amplify divisive agendas. A variety of interventions in the short and medium term by social media platforms, legacy media, civil society, educational institutions, and the state can disrupt these dynamics and mitigate social harm.

Public education

Political discussion on social media returns again and again to a relatively small set of problematic themes and discursive strategies that are used to insult, silence, and undermine.

Four themes in particular cross between different groups and situations and are used on both the far left and right: Nazi analogies and invocations of genocide; animal analogies; accusations of "selling out"; and xenophobic speech.

The pervasive nature of these themes creates opportunity for interventions. The goal cannot be to root out such discourse entirely, but to marginalise it and raise the costs of using it.

One potential approach is to develop educational strategies to sensitise the public as to why such discourse is deeply problematic and to describe the social costs of resorting to language of this kind.

Such interventions would not only explain the problematic dynamics associated with invoking Nazism (and animalisation, "selling out", and xenophobic speech) in everyday discourse, but would also offer guidance as to how to debate contentious issues in less inflammatory ways. A discussion of Nazi analogies could thus, for example, serve two purposes: explaining what happens when Nazism is invoked (i.e., how it drives conversation in particular directions) and introducing "guard rails" for public debate, i.e. proposing a set of standards for conversation in a pluralistic society.

Establishing and policing norms

The latter goal can also be advanced by mobilising those with moral stature to develop and promote a public set of standards for online citizenship. This could take the form of a voluntary social compact expected of all "good" online citizens. Here, the dynamic of silencing and cancelling can be put to advantage; those who egregiously transgress these standards could be called out, marginalised, and silenced for transgressing these norms.

While some groups on social media will resist public shaming – and indeed may profit from additional attention – public pressure on other groups may change their thinking about how to engage on social media. Outrageousness, and the attention that it generates, has become a lucrative currency on social media. Outrageousness currently does not come with sufficient political and social cost. While it is unlikely that the EFF will desist from this proven mechanism for drawing attention, the African National Congress (ANC), Democratic Alliance (DA), and others that try to position themselves as representative of mainstream "respectable" opinion may be swayable and shame-able. The goal would be to raise the cost of outrageousness by consistently calling out and shaming those who adopt outrageousness as a tactic. This could be done in various ways – for example, by publicly contrasting the online behaviour of parties (like the ANC and DA) with the values that they claim to uphold (see above about the "social compact regarding online behaviour"), calling them out for invoking Nazism, etc.

Creating savvy and sensitive social-media users

Educational interventions should inform the public about discursive strategies used on social media, that is, to make them savvy users who can recognise all the mechanisms – whataboutism, name-calling, silencing, distortion – that are used to derail and toxify discussion, and able to call it out when they see it.

This approach should include workshops aimed at school groups designed to highlight the potential harm of posting or sharing problematic content.

Targeting individuals

Anonymity and fake accounts mean that there is no social cost for individuals who engage in inflammatory online behaviour. The reward system is distorted: outrageousness draws attention and traffic. There is no easy way to address this. The mechanisms of constraint suggested above (for example, calling out) are less likely to work with anonymous individuals than with organisations, and may in fact be counterproductive (i.e., drawing more attention to the unacceptable online behaviour and encouraging more outrageousness).

The Lerato Pillay case, discussed at length in this report, may suggest a strategy. Once "Pillay" was unmasked, he paid a real price for his online actions. The problem, at the moment, is that this case was highly unusual: most anonymous online provocateurs can safely act with impunity. Social media platforms are unlikely to provide much help here. But what may work is generating more cases like the unmasking of Lerato Pillay and ensuring that they appear with regularity. Such a strategy would necessitate partnering with organisations able to do the kind of behind-the-scenes tracing required to "out" particular bigots. Over time, the effect may be exemplary: persuading the public that there is potential cost to their behaviour on social media and encouraging them to think twice before resorting to online bigotry.

Pressuring the platforms

Social media platforms in South Africa are more successful at removing content relating to antisemitism than that involving racism.

As discussed in the report, this reflects a variety of technical factors. But it also points to the efficacy of interest groups in pressuring the social media platforms to act. This approach is difficult, slow, and frustrating, but has borne fruit when it comes to antisemitism. It thus provides a template for mobilising other interest groups to do the same. There is an opportunity for encouraging and coordinating this mobilisation. Given the current attention to the malign actions of social-media platforms, they may be susceptible to pressure, given their concerns about the introduction of additional government regulation.

In the American context, social-media platforms have been driven to act on content hosted on their platforms when threats to advertising profits are clear and imminent. Civil society campaigns targeted at brand safety, as well as targeting organisations such as Google Play and Apple App Store, have proven effective.

Over time, social-media platforms can and must be persuaded to invest in content moderation. At present, social media platforms see content moderation as incidental to their business practices rather than as a core part of their business. Rather than relying on an outsourced and precarious labour force that is scaled up in relation to individual crises, an increased number of moderators with cultural and linguistic proficiencies should be seen as a cost of doing business. The provision of facilities and psychological support to content moderators should also be seen as a cost of doing business.

Working with legacy media

Some of the most vitriolic discourse we encountered occurred in response to Facebook posts by legacy media institutions; these institutions seemed to dedicate no time and effort to moderating the discussion generated within the comments on Facebook that these articles generated. The move to monetising content on social-media platforms should not be seen as an opportunity to neglect this duty of care to the public.

Legacy media can also play a decisive role in highlighting the current dynamics of social media usage in South Africa and in exposing and "calling out" those who use it for malign purposes. The *Daily Maverick's* close collaboration with the Centre for Analytics and Behaviour Change (CABC) and Digital Forensics Research Lab (DFRLAB) suggests some of benefits of taking social media seriously.

Legislation and regulation

Even as the impending implementation of the Digital Services Act in the European Union has met with criticism that it leans too far towards limiting free speech on social media, such forms of legislation will be the most effective means of ensuring social-media platforms timeously moderate content on their platforms and adopt transparent procedures for doing so.

In addition to regulating social media platforms more effectively, existing policies, including the National Action Plan and Strategy to Combat Racism (NAP), will bear fruit if properly implemented.

Similarly, the South African Human Rights Commission (SAHRC) will be more effective in combatting online hate speech if it adopts a more transparent process. Decisions made by the SAHRC relating to online hate speech are unavailable on its website. Such material offers opportunities to educate the public and to broaden public discussion and debate about how to counteract online hate.

Further research

Building datasets containing hate speech found in the South African context will facilitate algorithmic interventions. Groups such as the Alan Turing Institute’s hub for online hate research have been set up to collate and organise resources for research and policymaking on online hate. However, their datasets currently contain no indigenous African languages. This must be a priority in the South African context.

Introduction

On 14 September 1991, the National Peace Accord was signed in South Africa by representatives of 26 political parties, interest groups, and national and homeland governments. These parties committed to fostering the emergence of a multiparty democracy. The Accord was signed during a period of increasing tension and violence as South Africa negotiated the transition from white minority rule. Not intended to replace the rule of law, the Accord was designed to add to it by providing a forum for resolving political and community conflicts. National, regional, and local peace committees and special criminal courts were established. Limitations were placed on the activities of the security forces and police. Peace structures were developed across the country as part of a participatory process.[1] Although not a success everywhere, the National Peace Accord "provided a rickety way across the divide between apartheid and democracy",[2] in the process "building communication and political tolerance".[3] This was a crucial step in the project of developing the "rainbow nation".

Just over a month earlier, and with far less fanfare, a very different type of project was launched. On 6 August 1991, the World Wide Web went live. Its inventor, Tim Berners-Lee, described it as a "powerful global information system" designed with the philosophy that information should be freely available to anyone.[4] Others came to embrace this vision of the web as a space of unmediated expression and social connection.[5] The idealism and hope that twinned these two projects, born at roughly the same time, has since eroded.

[1] 'Our Constitution', https://ourconstitution.constitutionhill.org.za/the-national-peace-accord-npa/ (accessed 14 January 2022).

[2] Susan Collin Marks, *Watching the Wind: Conflict Resolution During South Africa's Transition to Democracy* (Washington: United States Institute of Peace, 2000), pp. 8–10.

[3] Padraig O'Malley, 'The National Peace Accord and its Structures' (Nelson Mandela Foundation), https://omalley.nelsonmandela.org/omalley/index.php/site/q/03lv02424/04lv03275/05lv03294/06lv03321.htm (accessed 14 January 2022).

[4] Martin Bryant, '20 years ago today, the World Wide Web opened to the public', thenextweb.com, https://thenextweb.com/news/20-years-ago-today-the-world-wide-web-opened-to-the-public (accessed 12 January 2022).

[5] For examples of the internet's ability to create new networked publics on social-media platforms, see Nathan Rambukkana (ed.), *Hashtag Publics: The Power and Politics of Discursive Networks* (New York: Peter Lang, 2015).

In the years since the launch of the World Wide Web, the utopian ideals of the internet have been undercut by the proliferation of obscene, violent, pornographic, illegal, abusive, and hateful content.[6] Social-media platforms have exemplified the potential of the internet to further the ideals of social connection and unmediated expression, but also its problems as those who wish to propagate hate have too found an opportunity for social connection and unmediated expression and have relished the anonymity that online activity often provides. As a result, social-media platforms have come under increasing pressure to modify their content-moderation processes (or lack thereof).

In South Africa, we have seen ample evidence of the power of social media for good and ill. Social media played a significant role in mobilising large-scale social movements such as #RhodesMustFall and #FeesMustFall. They have also been fingered as a contributing element and facilitating agent in the unrest that wracked Gauteng and KwaZulu-Natal in July 2021. The Centre for Analytics and Behavioural Change identified and reported 12 Twitter accounts as responsible for "repeatedly retweeting hashtags intended to incite an uprising".[7] In the month of July 2021, these hashtags generated a total of 1.29 million mentions and more than one million retweets.[8]

Social media has also become the locus for the expression of hate (as well as its exposure). Several episodes have become infamous, including Vicki Momberg's racial abuse of a black police officer, Penny Sparrow's epithets aimed at black beachgoers, and Velaphi Khumalo's response that whites should be "hacked and killed like Jews".[9] While each of these individuals was found guilty of hate speech, discussion on Twitter following these events using the hashtags #VickyMomberg and #Pennysparrow quickly degenerated into the "participatory reproduction of racism".[10] This in itself is a testament to the increasingly polarised nature of discourse relating to race and belonging in South Africa more broadly but also suggests that social media has played a role in furthering polarisation.

As the Centre for Analytics and Behavioural Change's work indicates, social-media platforms provide researchers with a large and readily available archive that provides access to everyday conversation and debate. Yet social-media posts are not an unmediated access point for understanding the thought processes of the public. Rather, discussion on social media is shaped by the demographics of those who use these platforms, as well as by the various political and commercial

[6] Sarah T Roberts, 'Digital detritus: 'Error' and the logic of opacity in social media content moderation' in *First Monday*, Vol. 23, No. 3-5 (March, 2018).

[7] Molebogang Mokoka, 'Meet the instigators: The Twitter accounts of the RET forces network that incited violence and demanded Zuma's release', *Daily Maverick*, 25 July 2021.

[8] Ibid.

[9] For an analysis of the networked nature of the #RhodesMustFall and #FeesMustFall movements, see Tanja Bosch, 'Twitter activism and youth in South Africa: The case of #RhodesMustFall', *Information, Community and Society* 20, no. 2 (2017).

[10] Allen Munoriyarwa, 'There ain't no rainbow in the 'rainbow nation' in Marta Pérez-Escolar and José Manuel Noguera-Vivo (eds.), *Hate Speech and Polarization in Participatory Society* (London & New York: Routledge, 2021), pp. 67–82.

considerations that drive content-moderation practices.[11] All of these elements are discussed in depth in this report.

The aim of this report is to understand how racist, xenophobic and antisemitic content has manifested on three of the most popular social-media platforms in South Africa: Facebook, Twitter and TikTok. To do this, the report has been split into four sections. Section One examines the demographics of social media usage in the South African context. Section Two provides a thematic analysis of racist, xenophobic and antisemitic content systematically collected from Facebook, Twitter and TikTok. Section Three – an extended appendix – focuses on the architecture of each platform, describes their content-moderation processes, and explains the specific cultures of use associated with Facebook, Twitter, and TikTok.[12] Section Four provides a description of the methodology of this study, particularly the approach used to extract and code material drawn from the social-media platforms.

The flashpoints

Our research focused on four flashpoints chosen to highlight the ways in which antisemitism, racism and xenophobia manifest on social-media platforms in the South African context since 2020. We avoided events that have already been subject to significant analyses and took place before 2020 (such as #VickyMomberg and #Pennysparrow). In each instance, we conducted a qualitative textual content analysis of a substantial sample of posts.

The Senekal Protests, October 2020

The first flashpoint that was analysed involved protests that erupted in the small town of Senekal, the magistracy of a rural farming district in the eastern Free State. These protests occurred following the brutal murder of Brendin Horner, a white farm manager, whose body was found on 2 October 2020 and was killed by a suspected stock thief or thieves. His murder quickly fed into a pre-existing discussion about the killing of white farmers as well as a perceived lack of government intervention when it came to 'farm murders'. AfriForum, a lobby group that mobilises around white Afrikaner interests, quickly labelled Horner's murder, which came to stand as an exemplar of farm murders in general, as "an act of terrorism".

On 6 October 2020, a group of farmers and community members protested outside the Senekal Magistrate Court as two black African men were brought before the court. These protests erupted into violence as some of the protestors attempted to force their way into the holding cells, damaged court property, and overturned and set alight a police vehicle.

[11] Tanja E Bosch, *Social Media and Everyday Life in South Africa* (London and New York, Routledge, 2021), at page 4.

[12] Ariadna Matamoros-Fernández, 'Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube' in *Information, Communication and Society*, Vol. 20, No. 6 (2017), pp. 930–946.

The next appearance of the accused in court on 16 October 2020 was accompanied by another set of protests by white farmers, as well as by AfriForum and the right-wing Afrikaner survivalist group Kommandokorps. There were also counter-protests organised by the Economic Freedom Fighters (EFF), a left-wing pan-Africanist political party created by former African National Congress Youth League president Julius Malema following his expulsion from the African Nationalist Congress (ANC). The EFF claimed to be there to help protect state property and openly sang 'Kill the Boer' (*Dubula ibhunu*), a South African struggle song that was deemed as hate speech in 2011 by the Equality Court.

The Brackenfell High School protests, November 2020

The second flashpoint flagged for analysis began with a series of protests organised by both the EFF and the Pan Africanist Congress of Azania (PAC) in Brackenfell, a largely white suburb of Cape Town. The protests were driven by allegations that a parent-organised function for Brackenfell High School students on 17 October 2020 was only open to whites, that the presence of two teachers at this function was an indication of school support for the event, and that this was symptomatic of systemic racism at the school (and by implication within the broader community).

An initial protest on 9 November 2020 saw EFF protesters faced by local residents (who were by-and-large white); some EFF members were assaulted. A much more substantial EFF-led protest soon followed with a crowd estimated at 2000; a white resident was assaulted after attempting to speak to EFF secretary general Marshall Dlamini. Members of the EFF and PAC are reported to have chanted "Shoot the Boer" and "one settler one bullet". Police used stun grenades, sprayed dye, and released teargas to disperse the predominantly black EFF protestors.

These two flashpoints, one in a rural area, one in an urban area, one featuring claims of systemic racism in South Africa and the other featuring claims of government apathy in the face of a supposed 'white genocide' (as some Twitter posts claimed), generated intense discussion on social media over the question of land and belonging.

Operation Dudula

While the Brackenfell High School protests occurred in white Afrikaans suburbia, Operation Dudula took us to Diepkloof. Located on the eastern edge of Soweto, Diepkloof was the site of an 'anti-illegal foreigner' march named 'Operation Dudula' on 16 June 2021. Hundreds of Soweto residents confronted suspected drug dealers and forcefully removed illegal occupants of a building. A second Operation Dudula march took place in Hillbrow on 19 February 2022 to demand the removal of foreign nationals; the march was accompanied by the singing of xenophobic songs. The protestors claimed that foreigners are responsible for taking jobs that should go to South Africans, for the high crime levels in the area, and for the sale of drugs.

Dudula is a Zulu term meaning 'to push out' or 'to drive away', and the moniker Operation Dudula now refers to a movement rather than an event, with the aim of removing foreign nationals from the country. The movement is closely aligned to the #PutSouthAfricaFirst movement. The Put South Africa First hashtag has a significant national reach and was one of the top ten trending hashtags in South Africa in 2020, receiving thousands of daily uses.[13] Operation Dudula, however, provided us with a more focused entry point into the issue of xenophobia in South Africa in an area with very different racial and linguistic features than Brackenfell and Senekal.

The 2021 Gaza Conflict

For our final flashpoint, we chose an international event which gained significant traction in local media. This conflict was triggered by an anticipated court decision around the eviction of Palestinian families in Sheikh Jarrah in Jerusalem. These protests snowballed into a series of larger protests, as well as rocket attacks by Hamas and Islamic Jihad into Israel and Israeli airstrikes targeting the Gaza Strip. We felt responses to these events would give us a sense of how social media in South Africa responded to a contentious international event.

Our Approach

A team of eight researchers with a range of language skills coded sample datasets for each flashpoint. The database they developed was then analysed in greater detail by three additional researchers. The methods used to gather and analyse the data are discussed in Section Four.

By analysing flashpoints drawn from across the country (and, in one case, beyond South Africa's borders) that traversed areas with different racial, linguistic, and economic features, that occurred in both urban and rural contexts and that mobilised groups ranging from the far left to the far right, we hope to provide a clear indication of the various ways that racism, xenophobia, and antisemitism manifest on social media in South Africa.

As we will see, one of the most striking features is how quickly much of the discussion on social media of contentious political events morphed into racism, fear-mongering, and othering. Again and again, racist tropes stabilised into instantly recognisable forms. The discussion became polarised along racial lines. Users were pressured to conform and chastised if they did not. Ad hominin attacks, vitriolic (and sometimes violent) language were normalised.

These dynamics, as well as a variety of themes that point to toxic trends on social media in South Africa, are discussed in detail in the pages that follow.

[13] 'Top hashtags in South Africa in 2020'. https://www.talkwalker.com/blog/social-media-stats-south-africa. For an analysis of #PutSouthAfricaFirst's online footprint, see Superlinear, 'Xenophobia, nationalism & populism: What's going on with #PutSouthAfricansFirst?'. https://www.superlinear.co.za/xenophobia-nationalism-populism-whats-going-on-with-putsouthafricansfirst/

1 Social Media Use in South Africa

Physical and digital divides

According to the World Bank Gini index, South Africa is currently the most unequal society in the world, with a Gini coefficient of 63% in 2014 (59.4% in 1994).[14] Tanja Bosch notes in her analysis of race and racism on Twitter in South Africa that "[i]t is against this current socio-economic context that we should read social media […] engagements on race and racism in South Africa".[15] South Africa's past manifests on social media through racist terms and stereotypes, but also in who has access to these platforms.

Since the 1990s, the number of people across the world with digital access has exploded. According to Statista, as of January 2021, there were 4.66 billion active internet users worldwide (59.5% of the global population). Of this total, 92.6% (4.32 billion) accessed the internet via mobile devices.[16] Of those who do not have access to digital connectivity, 96% live in developing countries.[17]

These figures are mirrored in the South African context, where 60.73% of South Africa's population was estimated to have access to the internet in 2021.[18] The uneven nature of this access can be seen in figures 1.1 and 1.2 and has become the subject of increased academic and political attention, particularly with the shift to online learning in South Africa during the Covid-19 pandemic. This digital divide has led to increasing calls (both globally and in South Africa) to see internet

[14] By comparison, the Gini coefficient of the United Kingdom was 34% in 2014. Gini Index (World Bank estimate) – South Africa. https://data.worldbank.org/indicator/SI.POV.GINI?locations=ZA&most_recent_year_desc=false

[15] Tanja E Bosch, *Social Media and Everyday Life in South Africa* (London and New York, Routledge, 2021), at page 136.

[16] 'Global digital population as of January 2021', *Statista* https://www.statista.com/statistics/617136/digital-population-worldwide/#:~:text=As%20of%20January%202021%20there,the%20internet%20via%20mobile%20devices (accessed 23 January 2021).

[17] Cecilia Rodriguez, 'Why a Third of the World, Nearly Three Billion People, Have Never Used the Internet', *Forbes*, 2 December 2021 https://www.forbes.com/sites/ceciliarodriguez/2021/12/02/why-a-third-of-the-world-nearly-three-billion-people-have-never-used-the-internet/?sh=2695c5c36a3f (accessed 21 January 2022).

[18] 'Internet user penetration in South Africa from 2017 to 2026', *Statista* https://www.statista.com/statistics/484933/internet-user-reach-south-africa/#:~:text=This%20statistic%20provides%20information%20on,to%2066.06%20percent%20in%202026 (accessed 23 January 2022).

access as fundamental for the exercise and enjoyment of the right to freedom of expression and opinion.[19]

While internet access remains uneven both globally and within South Africa, it is still clear that South African society is becoming increasingly digital. Much discussion, debate, and conversation now take place online.

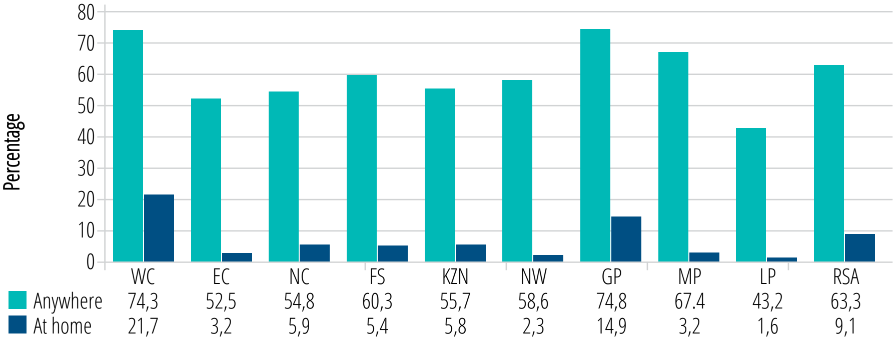


Figure 1.1: *Percentage of Households with access to the Internet at home, or for which at least one member has access to, or used the Internet by province, 2019.*[20]

Place where Internet is accessed	Rural / Urban status	Province (per cent)									
		WC	EC	NC	FS	KZN	NW	GP	MP	LP	RSA
At home	Metro	25,0	6,8	-	8,3	9,9	-	15,7	-	-	15,4
	Urban	16,1	3,5	5,8	4,3	6,5	3,9	9,4	6,2	5,6	7,2
	Rural	10,5	0,3	6,2	4,1	1,1	1,1	9,0	0,9	0,7	1,2
	Total	**21,7**	**3,2**	**5,9**	**5,4**	**5,8**	**2,3**	**14,9**	**3,2**	**1,6**	**9,1**
At work	Metro	28,1	24,9	-	15,3	28,7	-	29,1	-	-	28,0
	Urban	21,9	11,6	17,8	10,8	20,6	13,8	21,5	15,6	16,9	17,1
	Rural	9,8	5,1	10,4	7,1	4,7	4,8	5,5	5,2	5,0	5,2
	Total	**25,4**	**13,4**	**15,7**	**11,7**	**17,7**	**8,8**	**28,0**	**9,7**	**7,2**	**18,6**

[19] In the 2022 State of the Nation Address, the Communications Minister Khumbudzo Ntshaveni promised that government will provide 10GB of free data per month to every South African household and went on to state "[d]ata has become a new utility like water and electricity that our home needs". Khumbudzo Ntshaveni quoted in Unathi Nkanjeni, "Government will soon provide 10GB of free data to every household — here's how it'll work", *Sunday Times,* 17 February 2022. https://www.timeslive.co.za/news/south-africa/2022-02-17-government-will-soon-provide-10gb-of-free-data-to-every-household-heres-how-itll-work/

[20] Statistics South Africa, *General Household Survey, 2019*, at pages 51–52.

Place where Internet is accessed	Rural / Urban status	Province (per cent)									
		WC	EC	NC	FS	KZN	NW	GP	MP	LP	RSA
Using mobile devices	Metro	72,3	65,8	-	67,8	58,2	-	69,4	-	-	67,8
	Urban	54,2	48,6	53,8	54,4	58,0	66,2	63,3	74,8	50,5	59,5
	Rural	34,5	39,0	42,3	47,8	41,3	50,2	29,4	60,1	37,6	44,0
	Total	**65,2**	**50,3**	**50,5**	**57,7**	**51,7**	**57,2**	**68,4**	**66,4**	**40,4**	**58,7**
At Internet Cafes or educational facilities	Metro	18,9	15,8	-	13,2	11,3	-	17,6	-	-	16,6
	Urban	8,6	5,8	5,8	9,9	10,8	11,3	12,9	6,5	4,3	9,1
	Rural	1,1	3,1	1,8	6,9	2,2	6,2	0,0	2,7	1,7	2,9
	Total	**15,0**	**8,1**	**4,6**	**10,6**	**7,7**	**8,5**	**16,9**	**4,3**	**2,2**	**10,7**

Figure 1.2: *Households' access to the Internet by place of access, urban/rural status and province, 2019.*[21]

The rise of social-media platforms

Platforms are online sites and services that host, organise, and circulate content for users without producing a great deal of content themselves.[22] This can include search platforms (such as Google), social-media platforms (such as Facebook), and other industry platforms (such as Uber and Airbnb). We focus exclusively on social-media platforms. These companies now play key roles in mediating communications through public profiles, content feeds, private messaging, and other communication channels. They have become a means by which people communicate in their private lives as well as a fundamental part of the public sphere.[23]

Social-media platforms have transformed the global media landscape. A public that had until recently largely played a passive role as consumers of media content have now become producers and distributers of phenomenal quantities of user-generated content (see Figure 1.3). Social-media platforms encourage users to produce their own content, generally in the form of pictures, videos, and text.[24] This change from being users of content to producers of content has shaped how we socialise, organise, and communicate.[25]

[21] Ibid.

[22] See Tarleton Gillespie, *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media* (New Haven & London: Yale University Press, 2018).

[23] Jennifer Cobbe, 'Algorithmic Censorship by Social Platforms: Power and Resistance' in *Philosophy & Technology*, Vol. 34, No 7 (2021), pp. 739–766, at page 740.

[24] Dimitra Minitrakopolou, 'Social Media' in Laurie A Schintler & Connie L McNeely (eds.), *Encyclopedia of Big Data* (Springer International Publishing, 2022)**,** at page 186.

[25] Devan Rosen, 'Introduction: The Rise of a New Media Paradigm' in Devan Rosen (ed.), *The Social Media Debate: Unpacking the Social, Psychological, and Cultural Effects of Social Media* (New York: Routledge, 2022), at page 2.

Social-media platforms – some of which are multi-billion-dollar companies – offer not just access to information but also a constant audience of friends, family, and anonymous strangers. When logging in to Facebook, the platform asks, "What's on your mind?", Twitter similarly enquires "What's Happening?' and YouTube instructs "Broadcast Yourself", thus "encouraging content production and promoting users to believe that they themselves are the ones with the power to decide what can be posted and that they are spaces where they can express themselves however they see fit".[26]

These exhortations to produce new content also hint at the fact that the lifeblood of social-media platforms is user-generated content. It is the constant production of such content that keeps the userbase actively engaged.[27] In the process, the social-media platforms may also be tailoring content to fit user predilections and tastes by creating what are described as 'filter bubbles'.[28]

Social-media platforms offer us much more than an online space for communication and the sharing of user-generated content: they often act as marketplaces, payment systems, advertisers, gaming sites, and media distributors.[29] As a result, access to these platforms has become increasingly central to our ability to work, socialise and live our lives.[30] Figure 1.4 provides a visual example of these entanglements across platforms and services.

The struggles of old media news outlets have also opened a void for an information-hungry public that has been filled by social-media platforms that have now become both circulators of the news and shapers of it. The platforms are now powerful disseminators of information and act as the first-source news outlet for many users.[31] As a result:

> ***social media present the most far-reaching and comprehensive navigable social network that has ever existed, allowing users to locate and interact with others and affiliation groups faster and more globally than any previous communication and information technology. It can give voice to those who were previously silenced, while simultaneously aiding the spread of distorted information with dire consequences.***[32]

[26] GK Young, 'How much is too much: The difficulties of social media content moderation' in *Information & Communications Technology Law* (2021), pp. 1–16, at page 4.

[27] Sarah T Roberts, 'Digital detritus: 'Error' and the logic of opacity in social media content moderation ' in *First Monday*, Vol. 23, No. 3-5 (March, 2018).

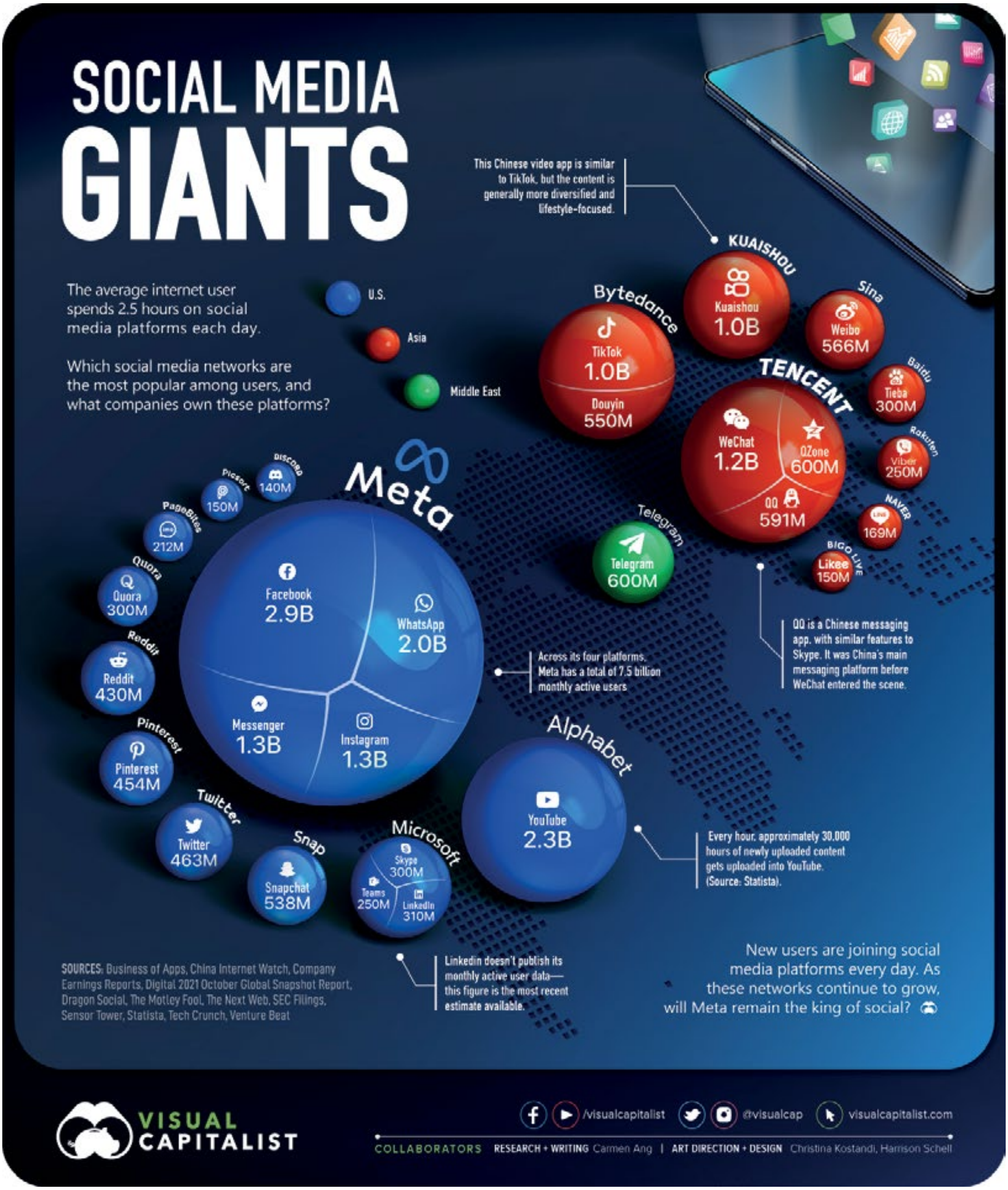
[28] Chinmayi Arun, Rebalancing Regulation of Speech: Hyper-Local Content on Global Web-Based Platforms (2018).

[29] Sarah Myers West in Tarleton Gillespie et al, 'Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates' in *Internet Policy Review*, Vol. 9, No. 4 (2020), pp. 1–30, at page 15.

[30] Ibid.

[31] Sarah T Roberts, 'Digital detritus: 'Error' and the logic of opacity in social media content moderation ' in *First Monday*, Vol. 23, No. 3-5 (March, 2018).

[32] Devan Rosen, 'Introduction: The Rise of a New Media Paradigm' in Devan Rosen (ed.), *The Social Media Debate: Unpacking the Social, Psychological, and Cultural Effects of Social Media* (New York:



The average internet user spends 2.5 hours on social media platforms each day.

Which social media networks are the most popular among users, and what companies own these platforms?

This Chinese video app is similar to TikTok, but the content is generally more diversified and lifestyle-focused.

Across its four platforms, Meta has a total of 7.5 billion monthly active users

QQ is a Chinese messaging app, with similar features to Skype. It was China's main messaging platform before WeChat entered the scene.

Every hour, approximately 30,000 hours of newly uploaded content gets uploaded into YouTube. (Source: Statista).

Linkedin doesn't publish its monthly active user data—this figure is the most recent estimate available.

New users are joining social media platforms every day. As these networks continue to grow, will Meta remain the king of social?

SOURCES: Business of Apps, China Internet Watch, Company Earnings Reports, Digital 2021 October Global Snapshot Report, Dragon Social, The Motley Fool, The Next Web, SEC Filings, Sensor Tower, Statista, Tech Crunch, Venture Beat

COLLABORATORS RESEARCH + WRITING Carmen Ang | ART DIRECTION + DESIGN Christina Kostandi, Harrison Schell

Figure 1.3.[33]

Routledge, 2022), at page 1.

[33] https://www.visualcapitalist.com/ranked-social-networks-worldwide-by-users/

User-generated content is social media's greatest currency but also one of its largest liabilities. The platforms have struggled with the proliferation of obscene, violent, pornographic, illegal, abusive and hateful content, as well as misinformation.[34] They have also been blamed for distorting socialisation patterns (for example, creating unrealistic expectations about body image), for flouting privacy, for fuelling internet addiction, and for affecting mental health. Various journalistic exposés have highlighted how platforms are aware of these problems but have often prioritised commercial expediency above active intervention.[35]

Although social networks are used by individual users, the platforms are also used by organised and semi-organised groups. Some extremist groups, for example, have used platforms to recruit adherents, propagandise, and sow discord.[36] Though platforms provide global reach, they also serve as spaces for very local engagements. It is in and through these very local engagements that rhetoric can slip into real-world action.

The demographics of social media usage in South Africa

Social media use in South Africa has grown substantially over the last two years, encouraged by strict Covid-19 lockdowns and a surge in the use of mobile telephones and apps.[37] Some social-media platforms in South Africa now boast user figures that constitute almost half of the South African population (Table 1.5). But who exactly are these users and how might their demographics shape the content that they see and produce? The latest *South African Social Media Landscape Report, 2021* (which focuses on social media usage in South Africa in 2020) provides the most detailed publicly available analysis of the South African social media landscape.[38]

[34] Sarah T Roberts, 'Digital detritus: 'Error' and the logic of opacity in social media content moderation' in *First Monday*, Vol. 23, No. 3-5 (March, 2018).

[35] See, for just one of many examples, the 'Facebook Files' exposés by the *Wall Street Journal* between September and December 2021.

[36] Natalie Alkiviadou, 'Hate speech on social media networks: Towards a regulatory framework?' in *Information & Communications Law*, Vol. 28, No. 1 (2019), pp. 19–35, at page 20.

[37] 'The biggest and most popular social media platforms in South Africa, including TikTok', BusinessTech, 1 July 2021, https://businesstech.co.za/news/internet/502583/the-biggest-and-most-popular-social-media-platforms-in-south-africa-including-tiktok/ (accessed 29 January, 2022).

[38] The report is intended to provide businesses with an overview of the social media landscape and to entice them to purchase more detailed information in order to maximise advertising interventions and brand management on social media. The Report is produced by Ornico (which describe itself as a "Brand Intelligence® solution built for marketing professionals first") and World Wide Worx (which describes itself as an "independent technology research and strategy organisation"). www.ornico.co.za; https://www.worldwideworx.com/what-we-do/

Platform	Users
Facebook	27 million
YouTube	24 million
Instagram	10 million
Twitter	9,3 million
TikTok	9 million
LinkedIn	8,4 million
SnapChat	7 million

Figure 1.4.[39]

The report focuses on users aged 15+ living in urban areas with 8000 or more inhabitants and is based on interviews with 24,000 respondents. This focus on the urban (and particularly metropolitan) landscape is likely due to the far lower internet (and social media) penetration in rural areas. The data produced is weighted to the Statistics South Africa population estimates and claims to provide an accurate representation of 27 million South Africans.[40] Despite the commercial aims of the report and the fact that it is skewed towards the urban context (and shows large variations across each half of 2020), it still provides some useful insights for delineating the broader terrain of social media usage in the South African context.

Language choice by social media users and its implications

English dominates social media in South Africa. According to Talkwalker (another consumer intelligence platform), 91.6% of all posts are in English, followed by Afrikaans, Russian, German, Portuguese, French, Spanish, Italian and Dutch, with 3.1% being listed as 'other'.[41] Our initial assumption when beginning this project was that these statistics were a product of how analytical software for social media coded language and that having annotators that were fluent in a range of South Africa's national languages would show very different language demographics. While we found almost no indication of any posts in Russian, German, Portuguese, French, Spanish and Italian in our datasets, English still completely dominated, followed by Afrikaans, with the notable exception of the dataset for Operation Dudula, on which more will be said below.

Of the posts extracted from Twitter relating to the Brackenfell protests, 2.3% contained isiXhosa, 1.7% contained isiZulu and 1.1% were made up of tweets containing Sepedi, Sesotho, Setswana, and Tshivendi. Two and a half percent were exclusively in Afrikaans (rising to 4.36% of tweets that contained some Afrikaans) while the remainder of the posts were in English. Very different trends, however,

[39] 'The biggest and most popular social media platforms in South Africa, including TikTok', BusinessTech, 1 July 2021, https://businesstech.co.za/news/internet/502583/the-biggest-and-most-popular-social-media-platforms-in-south-africa-including-tiktok/ (accessed 29 January, 2022).

[40] South African Social Media Landscape Report, 2021 (Ornico and World Wide Worx, 2021), at page 62.

[41] 'Social media statistics and usage in South Africa', https://www.talkwalker.com/blog/social-media-stats-south-africa

emerged in our Facebook dataset relating to the Brackenfell protests. Here, 73.4% of the posts coded by our annotators were in English and 20.8% in Afrikaans (which goes up to 24.4% when we include posts that were in both English and Afrikaans). Only 2.2% of coded posts contained other national languages (isiXhosa, isiZulu, Setswana and Xitsonga).

The material extracted relating to the Senekal protests showed similar results. When it came to Twitter, 95.3% of all coded tweets were in English and 4.1% contained Afrikaans, while only 0.7% contained other languages. Meanwhile, the coded posts extracted from Facebook, although still dominated by English, contained 27% of posts coded as being in Afrikaans, with only 2.4% coded as containing isiXhosa.

While the number of Afrikaans posts may, to some extent, be a reflection of the topics chosen for analysis (Brackenfell is a predominantly Afrikaans-speaking northern suburb of Cape Town and the Senekal protests saw a particularly vigorous response by farmers who are also predominantly Afrikaans), this data highlights the ways that platform architecture shapes online discourse. Twitter does not support any of South Africa's eleven official languages other than English; the same is true of TikTok. Facebook, on the other hand, also provides support for Afrikaans (including a translation tool). This may in part explain why Facebook has the largest number of white users of the three platforms under consideration (see Figure 1.9 below).

In contrast, there was little to no Afrikaans at all on the datasets relating to Operation Dudula.[42] Posts on Twitter were predominantly in English, which formed 81% of the dataset. However, 12.5% of tweets were coded as isiZulu and 2.1% in Sepedi. The large increase in the number of posts containing isiZulu here is unsurprising when one considers that it is the language most spoken by individuals in households across Gauteng and that the epicentre of the Operation Dudula movement was Diepkloof. Interestingly, this pattern is not reflected in posts extracted from Facebook.[43]

The dominance of English in these datasets, while not surprising, still requires further explanation, given how few South Africans speak English inside and outside the home (see Figure 1.5). A small study of language use on social media in Limpopo provides a likely explanation. As a result of the lack of recognition for African languages on social-media platforms, users found it "difficult to reflect on their cultures" when using social media as a means of communication.[44] English allowed these users to connect with people from different countries and social backgrounds. Using English saved time, as other users would otherwise often ask for translations of what they had written. When words are written in an

[42] 1.9% of extracted coded tweets and 0% of coded posts extracted from Facebook were in Afrikaans.

[43] The remaining posts were coded as 'other' as they often consisted only of an image or were made up of dead links where the post could not be coded. Of the extracted coded posts, 76.8% were listed as being in English and only 8.4% of coded posts were listed as containing indigenous languages (only 2.1% of these being in isiZulu).

[44] Edgar Malatji and Carol Lesame, 'The use of South African languages by youth on social media: The case of Limpopo Province' in *Communicare: Journal for Communication Sciences in Southern Africa*, Vol. 38, No. 1 (2019), pp. 76–95, at page 78.

African language, the platforms typically indicate that the word is unknown or not recognisable.[45]

	Black African		Coloured		Indian / Asian		White		South Africa	
	Inside	Outside	Inside	Outside	Inside	Outside	Inside	Outside	Inside	Outside
Afrikaans	0,9	1,0	77,4	68,8	1,3	1,5	61,2	37,2	**12,2**	**9,7**
English	1,6	8,6	20,1	28,3	92,1	95,8	36,3	61,0	**8,1**	**16,6**
Isindebele	1,9	1,6	0,0	0,0	0,3	0,2	0,3	0,1	**1,6**	**1,3**
IsiXhosa	18,2	15,6	1,1	1,3	0,4	0,0	0,1	0,1	**14,8**	**12,8**
IsiZulu	31,1	30,8	0,3	0,3	0,9	1,0	0,5	0,5	**25,3**	**25,1**
Khoi, Nama and San Languages	0,1	0,1	0,0	0,0	0,0	0,0	0,0	0,0	**0,1**	**0,1**
Sepedi	12,4	12,0	0,3	0,2	0,5	0,2	0,1	0,3	**10,1**	**9,7**
Sesotho	9,7	9,6	0,1	0,2	0,1	0,3	0,0	0,1	**7,9**	**7,8**
Setswana	11,1	11,5	0,7	0,8	0,2	0,2	0,4	0,4	**9,1**	**9,4**
Sign Language	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	**0,0**	**0,0**
SiSwati	3,5	3,2	0,0	0,0	0,0	0,0	0,0	0,0	**2,8**	**2,6**
Tshivenda	3,1	2,7	0,0	0,0	0,2	0,0	0,0	0,0	**2,5**	**2,2**
Xitsonga	4,4	2,9	0,0	0,1	0,1	0,1	0,0	0,0	**3,6**	**2,4**
Other	2,1	0,5	0,1	0,0	4,0	0,7	1,1	0,5	**1,9**	**0,5**
Total Percentage	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
Total (Thousands)	46 307	46 135	4 961	4 930	1 430	1 426	4 442	4 420	57 143	56 914

Figure 1.5: *Percentage of languages spoken by household members inside and outside household by population, 2018.*[46]

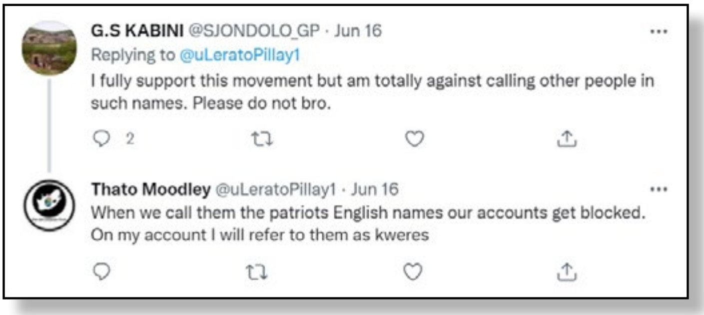
[45] Edgar Malatji and Carol Lesame, 'The use of South African languages by youth on social media: The case of Limpopo Province' in *Communicare: Journal for Communication Sciences in Southern Africa*, Vol. 38, No. 1 (2019), pp. 76–95, at page 88.

[46] 'These are the most-spoken languages in South Africa in 2019' in BusinessTech, 1 June 2019. https://businesstech.co.za/news/business/319760/these-are-the-most-spoken-languages-in-south-africa-in-2019/

The use of English by second-language speakers produced problems of its own. Our datasets showed numerous examples of users being ridiculed for poor use of English, with this often used as a means of dismissing their viewpoints and advancing the claim that schooling had deteriorated under black majority rule.

The failure of social-media platforms to support South African languages creates another problem. At present, there are few signs that languages such as isiZulu are consistently flagged by algorithmic content-moderation processes or understood by commercial content moderators. As we will see, this provides an opportunity for those intent on disseminating hateful content in these languages.

The increased use of isiZulu on Twitter by active supporters of Operation Dudula may in part reflect a recognition that they enjoy greater latitude when using this language. This strategy was explicitly promoted by @uLeratoPillay1, who we will discuss in greater detail later in this report. In response to a user who asks them to not use the term *kwerekwere* (derogatory slang that refers to black foreigners in South Africa), @uLeratoPillay1 responds: “When we call them the patriots English names our accounts get blocked. On my account I will refer to them as kweres“ (Figure 1.6). Use of this term is thus clearly a strategy to avoid content moderation.



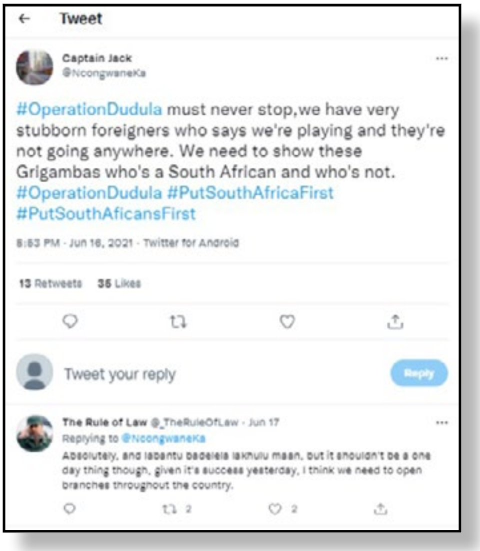
G.S KABINI @SJONDOLO_GP · Jun 16
Replying to @uLeratoPillay1
I fully support this movement but am totally against calling other people in such names. Please do not bro.

Thato Moodley @uLeratoPillay1 · Jun 16
When we call them the patriots English names our accounts get blocked. On my account I will refer to them as kweres

Figure 1.6

Another example of the use of language to avoid detection can be seen in Figure 1.7, a short interaction between two users where one refers to foreigners as “stubborn“ and as “Grigambas“ (a contemptuous term used to refer to inhuman things).[47] The response to this tweet is, “Absolutely, and labantu badelela lakhulu maan [these people are very contemptible], but it shouldn’t be a one day thing though, given it’s success yesterday, I think we need to open branches throughout the country“. Note here the move away from English when referring to foreigners as ‘contemptible’.

[47] Godfrey Mwakikagile, *Africa and its People* (Pretoria: New Africa Press, 2008), at page 268. See also Kenneth Tafira, ‘Is xenophobia racism?’ in *Anthropology Southern Africa*, Vol. 34. No. 3-4 (2011), pp. 114-121.



Captain Jack
@NcongwaneKa

#OperationDudula must never stop,we have very stubborn foreigners who says we're playing and they're not going anywhere. We need to show these Grigambas who's a South African and who's not. #OperationDudula #PutSouthAfricaFirst #PutSouthAficansFirst

5:53 PM · Jun 16, 2021 · Twitter for Android

13 Retweets 35 Likes

The Rule of Law @_TheRuleOfLaw · Jun 17
Replying to @NcongwaneKa
Absolutely, and labantu badelela lakhulu maan, but it shouldn't be a one day thing though, given it's success yesterday, I think we need to open branches throughout the country.

2 2

Figure 1.7

The racial and social demographics of social media users in South Africa

Unsurprisingly, whites are over-represented on social-media platforms in comparison to the racial demographics of South Africa as a whole (see figures 1.9 and 1.10). It was only over the course of 2020 that the number of black and white users on Twitter reached relative parity. This increase in black users on Twitter may be the result of a growing perception that Twitter is a more representative space with the rise of "Black Twitter". This may have also been facilitated by the introduction of new features such as Spaces.[48] The relative absence of black users on TikTok may help to explain why the large majority of material relating to the flashpoints on this platform was hostile to the EFF.

[48] *South African Social Media Landscape Report, 2021* (Ornico and World Wide Worx, 2021), at page 64.

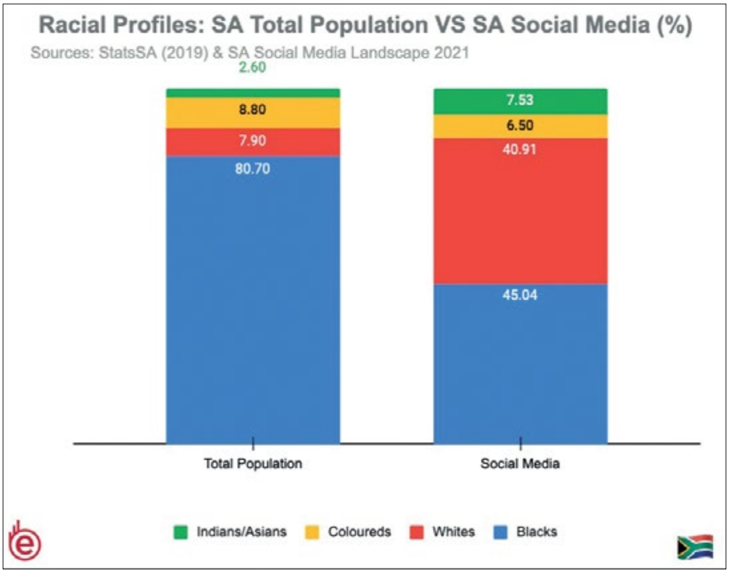


Figure 1.8.[49]

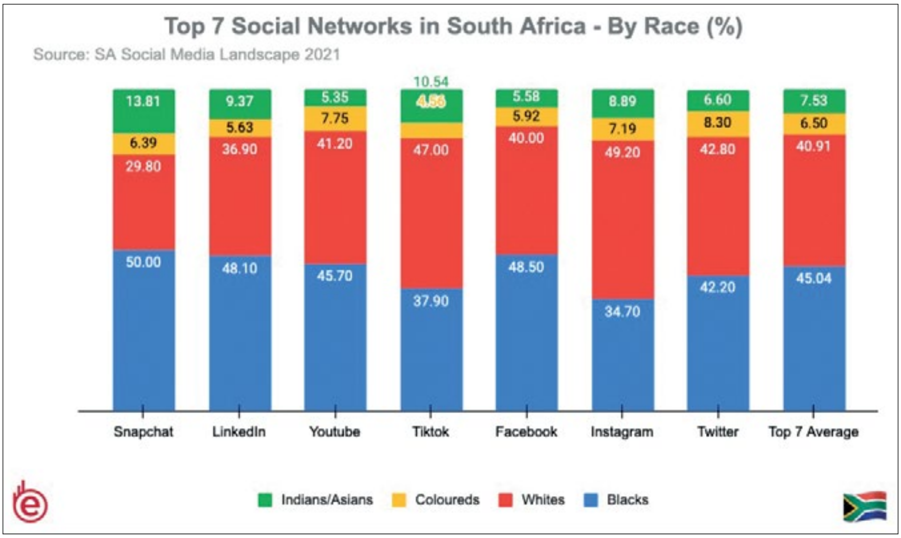


Figure 1.9.[50]

[49] Graphic compiled by Willy Seyama based on statistics taken from the South African Social Media Landscape Report, 2021. https://enitiate.solutions/social-media-landscape-2021-from-data-to-insights-for-brands/

[50] Ibid.

Discussion on social media is further skewed by the fact that the majority of users (across all racial groups) are made up of individuals from higher income bands. A third of South Africa's urban social media users come from the four highest income brackets, while only 12% are made up of individuals from the bottom four wage brackets (see Figure 1.10). This may be exacerbated by high data costs in South Africa.[51]

Social media in South Africa is thus heavily skewed in terms of both race and wealth. This has a variety of implications for the nature and tenor of discussions of contentious issues. Paradoxically, it may in part explain how and why the EFF so often sits at the centre of online debate. A significant proportion of social media users see the EFF as being antithetical to their interests, leading to greater antagonism towards the party and its supporters. The EFF, for example, completely dominates the polarised and polarising discussion relating to the protests in Brackenfell and Senekal. As we will see later, the EFF is itself skilled at driving discussion online.

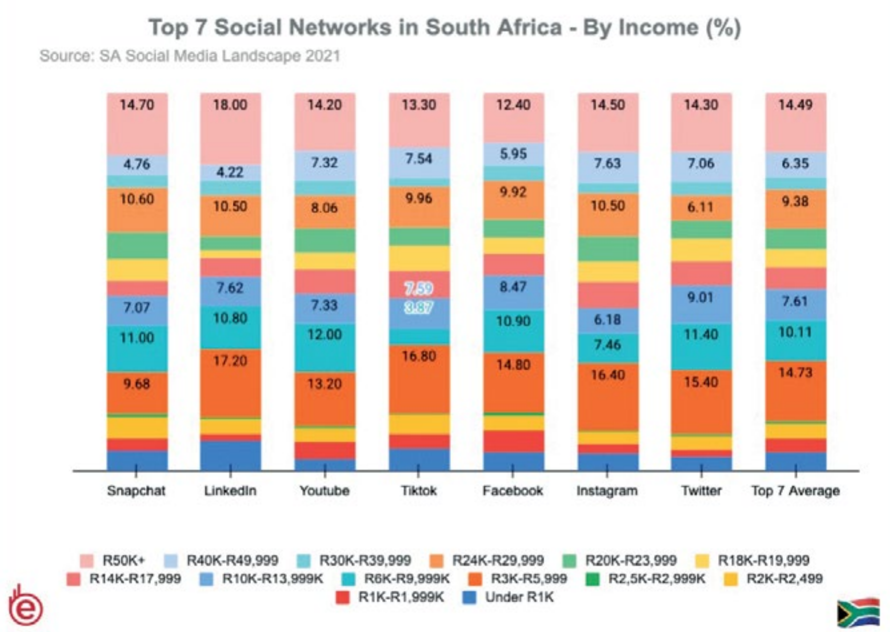
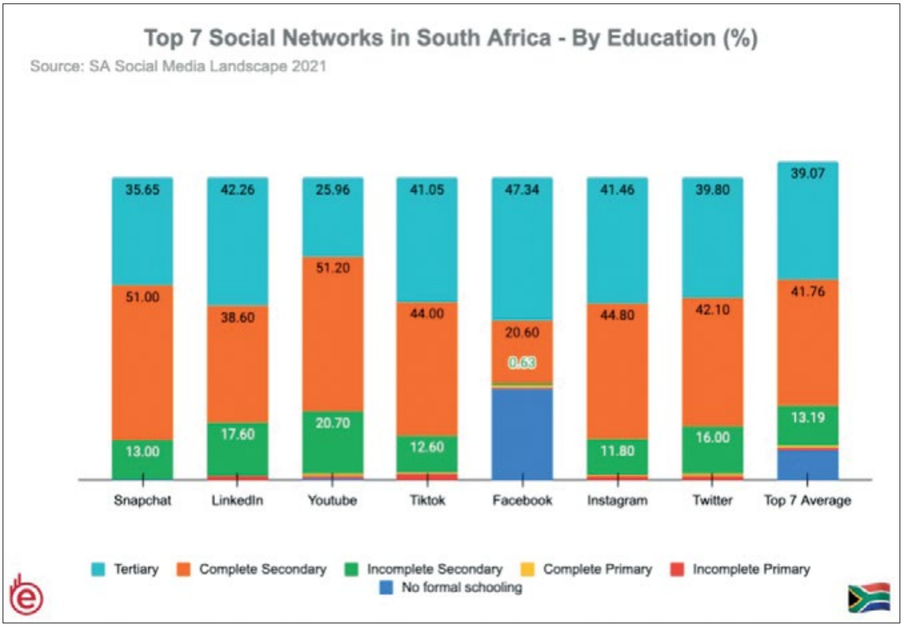


Figure 1.10.[52]

[51] https://www.cable.co.uk/mobiles/worldwide-data-pricing/. It is unclear whether these high data costs have been offset by the introduction of Facebook's 'Free Basics' in South Africa. The only analysis of Free Basics and its usage in South Africa is a study of its use by 35 individuals in the context of tertiary educational facilities in urban Cape Town. See Julianne Romanosky and Marshini Chetty, 'Understanding the Use and Impact of the Zero-Rated Free Basics Platform in South Africa' in CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Paper 192 (April, 2018).

[52] Graphic compiled by Willy Seyama based on statistics taken from the South African Social Media Landscape Report, 2021. https://enitiate.solutions/social-media-landscape-2021-from-data-to-insights-for-brands/


Top 7 Social Networks in South Africa - By Education (%)

Source: SA Social Media Landscape 2021

Figure 1.11.[53]

The age demographics of social media users in South Africa

Social media use is dominated by those under the age of 34. Young people may be more inclined to hold radical political views and to voice these in less modulated ways on social media.[54] Although much of the narrative around TikTok suggests that it has come to dominate the younger demographic, that does not seem to be the case in South Africa. The 2021 *South African Social Media Landscape Report* indicates that Facebook has more users in the 15–24-year-old demographic; both were far ahead of Twitter in this demographic (see Figure 1.12). The report itself suggests that a large proportion of TikTok users are likely those in secondary school or who have just finished school.[55] This, along with TikTok's concerted attempts to 'keep the app fun', may explain why there was much less content relating to the Brackenfell and Senekal protests and Operation Dudula. This may also be a result of TikTok's content-moderation policies (discussed in Section Three).

[53] Ibid.

[54] See 'Study shows young South Africans have no faith in democracy and politicians', *The Conversation*, 11 June 2019. https://theconversation.com/study-shows-young-south-africans-have-no-faith-in-democracy-and-politicians-118404 and Collette Schulz-Herzenberg, 'The South African non-voter: An analysis' in The Midpoint – Paper Series N° 2/2020. https://www.kas.de/documents/261596/10543300/The+South+African+non-voter+-+An+analysis.pdf/acc19fbd-bd6d-9190-f026-8d311078b670?version=1.0&t=1608

[55] *South African Social Media Landscape Report, 2021* (Ornico and World Wide Worx, 2021), at page 92.

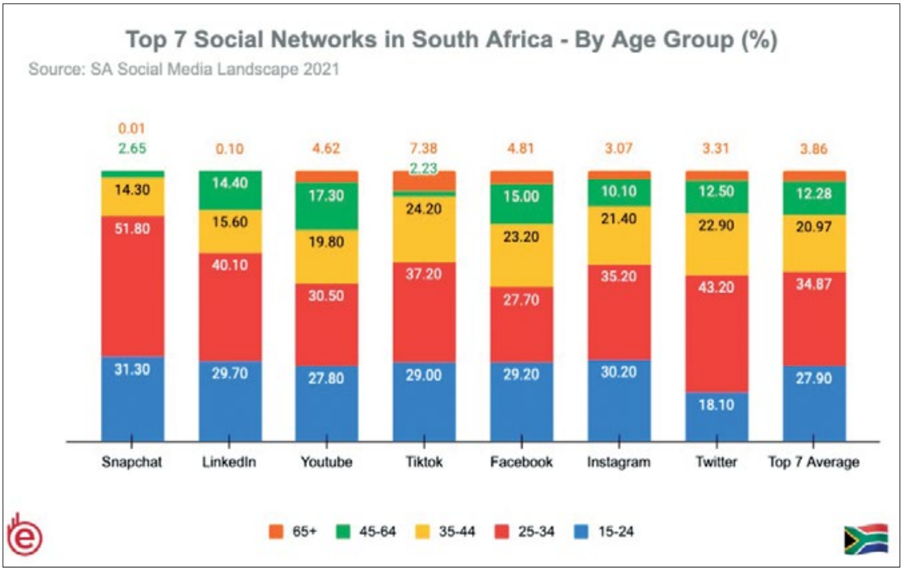


Figure 1.12: *Top seven social networks in South Africa by age group.*[56]

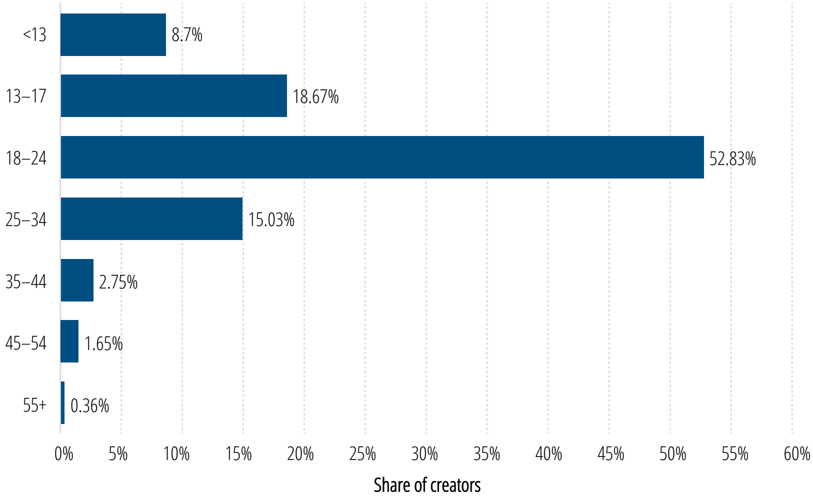


Figure 1.13: *Distribution of TikTok creators worldwide as of August 2021, by age group.*[57]

[56] Graphic compiled by Willy Seyama based on statistics taken from the *South African Social Media Landscape Report*, 2021. https://enitiate.solutions/social-media-landscape-2021-from-data-to-insights-for-brands/

[57] 'Distribution of TikTok creators worldwide as of August 2021, by age group', *Statista*. https://www.statista.com/statistics/1257721/tiktok-creators-by-age-worldwide/

Each of the features described above – the ways that social media usage reflects particular age, income, and racial patterns in South Africa – must be kept in mind when reading the next section of this study. As we will see, the presence of particular recurrent themes and motifs, as well as the dynamics of discussion, are in part a product of these patterns.

2 Flashpoints, dynamics and social media strategies

Section One of this report traced the landscape of social media use in South Africa. In Section Two, we identify and analyse key features and themes relating to each flashpoint, as well as identify the dynamics of online political discussion and the tools and strategies that users deploy to shape conversations.

This section builds on a close examination of a mass of content that was categorised and evaluated by the research team. As will be described in detail in Section Three, content on social media is ever-shifting, not only because of changes to what is written and posted, but also because of changes to community guidelines and the removal of accounts (and all their posts) if they are flagged or are found post facto to have violated the community guidelines of the social-media platform. Content relating to each of the flashpoints, in other words, changes over time due to deletions and additions.

Given the issue of fake accounts, we drew information from user profiles with caution, typically adding the disclaimer 'ostensibly' when describing the racial background of users. This is not to say that all the accounts that we describe are fake but that we should avoid taking the profiles at face value.

Despite the differences between the flashpoints, the most striking feature common to discussions of each on social media was how quickly conversations became polarised and shifted away from the events themselves into broader debates regarding the South African political scene. Some of those who participated in these discussions mobilised a range of racist tropes that are discussed in this section. As we will see, these tropes were often interrelated and appeared in multiple configurations. We also describe some of the more concerning dynamics of online discussion, allowing the reader to get a sense of how the white far right and the black radical left feed into and off one another and strengthen each other.

Part 2.1 examines the adroitness of the EFF at driving and dominating discussion on social media. Part 2.2 describes how the epithet of "race traitor" is deployed to demarcate and police political and racial boundaries on Twitter, Facebook, and TikTok. Part 2.3 explains how the far right presents whites as victims of the new South Africa and portrays the latter as dystopian. Part 2.4 demonstrates how the claim of "white genocide" builds international support for the far right in South Africa. Part 2.5 turns to the 2021 Gaza conflict in order to describe how content moderation and local concerns shape conversation on social media. Part 2.6 looks at the curious pull that Hitler, Nazi, and Holocaust analogies exercise on social media in South Africa. And 2.7 explores how fake accounts and other tools are deployed to translate social media activism into real-world action through the case study of Operation Dudula.

2.1 The EFF and the Dynamics of Social Media in South Africa

In 2012, the ANC found Julius Malema guilty of "sowing divisions within the ruling party" after he unfavourably compared the leadership of President Jacob Zuma to that of his predecessor Thabo Mbeki.[59] After his departure from the ANC he established his own political party, the Economic Freedom Fighters, in 2013. The EFF positioned itself as a radical, leftist, anti-capitalist and anti-imperialist movement focusing on issues of economic inequality. Despite being substantially smaller than the ANC and the official opposition Democratic Alliance (a broadly centrist political party), it has proven more successful than either of those two parties at driving debate on social media.[60]

How has it done so? Central to the party's political repertoire is its use of various media platforms to propagate its position on the expropriation of land and economic transformation.[61] These interventions are often framed provocatively and ambiguously. A good example of the latter is the claim made in multiple posts that "All we want is our land". Left unsaid is who this 'we' is, what land is being referred to, and how the process should take place. Such vague but deliberately provocative statements tend to stimulate a 'frenzy' on social media as they are shared and commented upon, with each new iteration of the post driving further engagement.[62]

Such posts were also noticeable for the level of resentment and anger they elicited.[63] In the content we collected, responses to the EFF often took the form of portraying the party as 'terrorists' and 'hooligans'. This occurred with such regularity that we soon stopped coding these posts. The image macro seen in Figure 2.1.1 below appeared on multiple occasions. So too did comments that played on various discourses of dysfunction (see, for example, figures 2.1.2 to 2.1.5). Here and elsewhere, criticism of the EFF was often barely concealed racism rather than commentary about the actions of the party. The framing of the EFF as terrorists allows for the party to be portrayed as enemies of the state and worthy targets of violence (see Figure 2.1.6).

[59] Nickolaus Bauer, 'Out! ANC upholds Julius Malema's expulsion', *Mail & Guardian*, 24 April 2012. https://mg.co.za/article/2012-04-24-malema-expelled/

[60] 'Julius Malema - South Africa's radical agenda setter', BBC News, 30 April 2019, https://www.bbc.com/news/world-asia-pacific-14718226

[61] Denél Chetty, #SCANDAL: An exploration of social media in light of René Girard's mimetic theory (Masters Thesis: University of Pretoria, 2020), at page 36.

[62] Ibid., page 41.

[63] Ibid., page 42.

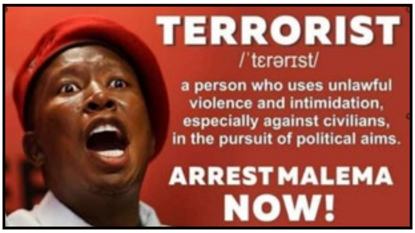
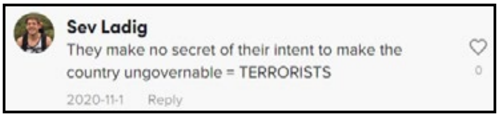

TERRORIST
/ˈtɛrərɪst/
a person who uses unlawful violence and intimidation, especially against civilians, in the pursuit of political aims.
ARREST MALEMA
NOW!

Figure 2.1.1

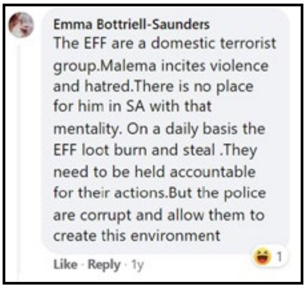

Sev Ladig

They make no secret of their intent to make the country ungovernable = TERRORISTS

2020-11-1 Reply

0

Figure 2.1.2

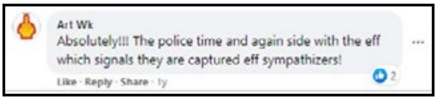

Emma Bottriell-Saunders

The EFF are a domestic terrorist group.Malema incites violence and hatred.There is no place for him in SA with that mentality. On a daily basis the EFF loot burn and steal .They need to be held accountable for their actions.But the police are corrupt and allow them to create this environment

Like · Reply · 1y

1

Figure 2.1.3

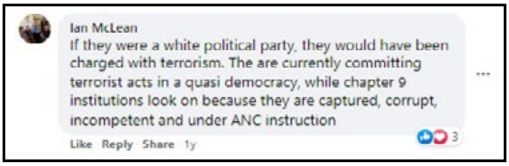

Art Wk

Absolutely!!! The police time and again side with the eff which signals they are captured eff sympathizers!

Like · Reply · Share · 1y

2

Figure 2.1.4

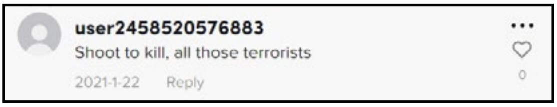

Ian McLean

If they were a white political party, they would have been charged with terrorism. The are currently committing terrorist acts in a quasi democracy, while chapter 9 institutions look on because they are captured, corrupt, incompetent and under ANC instruction

Like · Reply · Share · 1y

3

Figure 2.1.5



user2458520576883

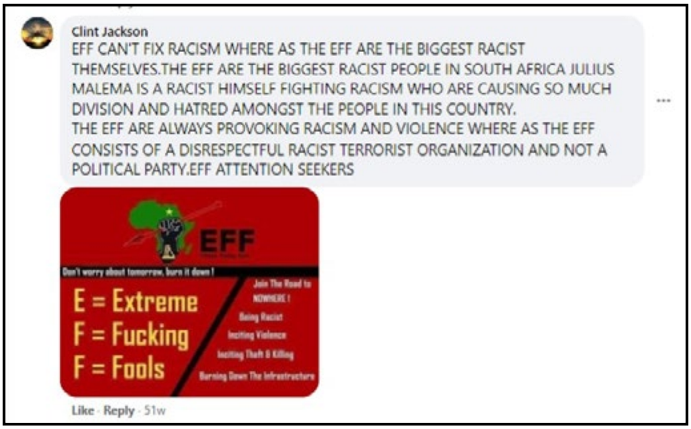
Shoot to kill, all those terrorists

2021-1-22 Reply

0

Figure 2.1.6

Similar sentiments appeared in content relating to the Brackenfell protests on Twitter and Facebook (see figures 2.1.7 and 2.1.8). Here we not only see the attempt to delegitimise and silence the EFF and its supporters by portraying them as terrorists and hooligans, but also the insinuation that EFF protestors are jobless individuals who mindlessly assemble as a paid 'rent-a-crowd'. The latter view permeated the datasets (see, for example, figures 2.1.9 to 2.1.11). Often paired with this is the claim that those who are unemployed are jobless because of their own laziness rather than because of systemic inequalities.



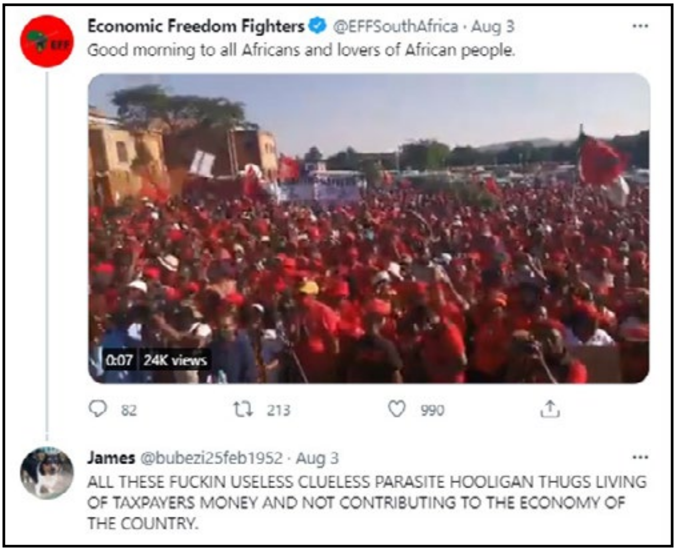
Clint Jackson

EFF CAN'T FIX RACISM WHERE AS THE EFF ARE THE BIGGEST RACIST THEMSELVES.THE EFF ARE THE BIGGEST RACIST PEOPLE IN SOUTH AFRICA JULIUS MALEMA IS A RACIST HIMSELF FIGHTING RACISM WHO ARE CAUSING SO MUCH DIVISION AND HATRED AMONGST THE PEOPLE IN THIS COUNTRY.
THE EFF ARE ALWAYS PROVOKING RACISM AND VIOLENCE WHERE AS THE EFF CONSISTS OF A DISRESPECTFUL RACIST TERRORIST ORGANIZATION AND NOT A POLITICAL PARTY.EFF ATTENTION SEEKERS

EFF
Don't worry about tomorrow, burn it down !
E = Extreme
F = Fucking
F = Fools
Join The Road to NOWHERE !
Being Racist
Inciting Violence
Inciting Theft & Killing
Burning Down The Infrastructure

Like · Reply · 51w

Figure 2.1.7



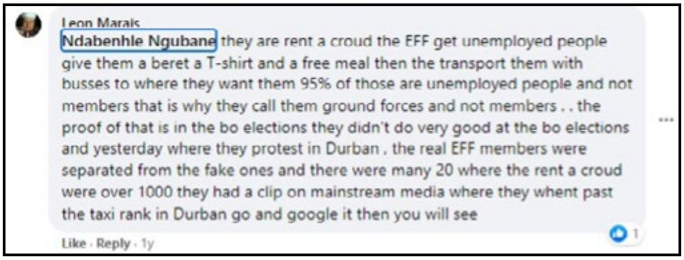
Economic Freedom Fighters @EFFSouthAfrica · Aug 3

Good morning to all Africans and lovers of African people.

James @bubezi25feb1952 · Aug 3

ALL THESE FUCKIN USELESS CLUELESS PARASITE HOOLIGAN THUGS LIVING OF TAXPAYERS MONEY AND NOT CONTRIBUTING TO THE ECONOMY OF THE COUNTRY.

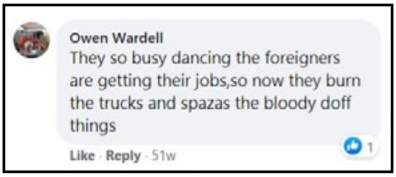
Figure 2.1.8



Leon Marais

Ndabenhle Ngubane they are rent a croud the EFF get unemployed people give them a beret a T-shirt and a free meal then the transport them with busses to where they want them 95% of those are unemployed people and not members that is why they call them ground forces and not members . . the proof of that is in the bo elections they didn't do very good at the bo elections and yesterday where they protest in Durban . the real EFF members were separated from the fake ones and there were many 20 where the rent a croud were over 1000 they had a clip on mainstream media where they whent past the taxi rank in Durban go and google it then you will see

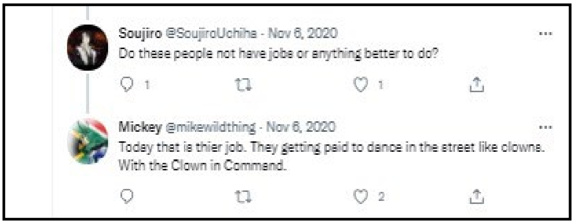
Figure 2.1.9



Owen Wardell

They so busy dancing the foreigners are getting their jobs,so now they burn the trucks and spazas the bloody doff things

Figure 2.1.10

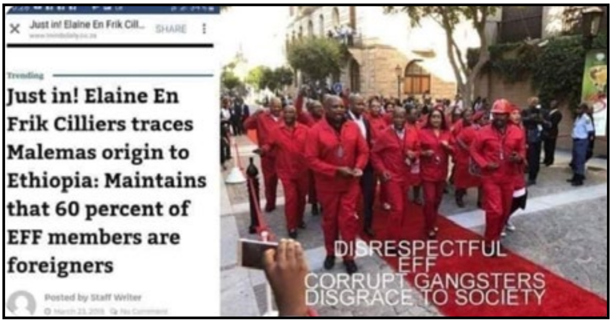


Soujiro @SoujiroUchiha · Nov 6, 2020
Do these people not have jobs or anything better to do?

Mickey @mikewildthing · Nov 6, 2020
Today that is thier job. They getting paid to dance in the street like clowns.
With the Clown in Command.

Figure 2.1.11

The assertions that the EFF is a terrorist organisation *and* a political party that does not enjoy genuine popular support are often accompanied by another strategy of silencing: the claim that the EFF and its supporters are foreigners. Many examples of this can be seen in relation to the Brackenfell protests. In response to the party's declaration on its Facebook page that "We will defend our land with our blood. This, our land we will defend with our bodies", a user simply posts a screenshot of an article claiming that Malema is a foreigner and that 60% of EFF members are foreigners (see Figure 2.1.12).[64] While this may seem to be an obvious example of 'fake news', the idea that Malema and his supporters are foreigners gained a great deal of traction. Others rushed to echo this claim (see, for example, figures 2.1.13 and 2.1.14), demonstrating the power of wilful belief and the no-holds-barred nature of the discussion. Here we see the intermeshing of xenophobia and racism, as well as fear-mongering, as detractors warned that the EFF is seeking to stoke a race war.



Just in! Elaine En Frik Cill...

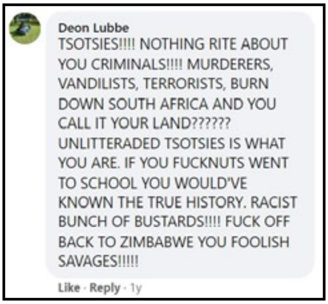
Trending

Just in! Elaine En Frik Cilliers traces Malemas origin to Ethiopia: Maintains that 60 percent of EFF members are foreigners

Posted by Staff Writer

Figure 2.1.12

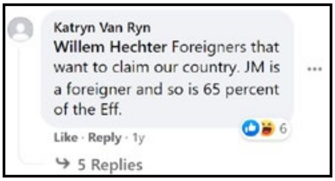
[64] https://www.trendsdaily.co.za/trending/just-in-elaine-en-frik-cilliers-traces-malemas-origin-to-ethiopia-maintains-that-60-percent-of-eff-members-are-foreigners/ (accessed 21 January 2022).



Deon Lubbe
TSOTSIES!!!! NOTHING RITE ABOUT YOU CRIMINALS!!!! MURDERERS, VANDILISTS, TERRORISTS, BURN DOWN SOUTH AFRICA AND YOU CALL IT YOUR LAND?????? UNLITTERADED TSOTSIES IS WHAT YOU ARE. IF YOU FUCKNUTS WENT TO SCHOOL YOU WOULD'VE KNOWN THE TRUE HISTORY. RACIST BUNCH OF BUSTARDS!!!! FUCK OFF BACK TO ZIMBABWE YOU FOOLISH SAVAGES!!!!!

Like · Reply · 1y

Figure 2.1.13



Katryn Van Ryn
Willem Hechter Foreigners that want to claim our country. JM is a foreigner and so is 65 percent of the Eff.

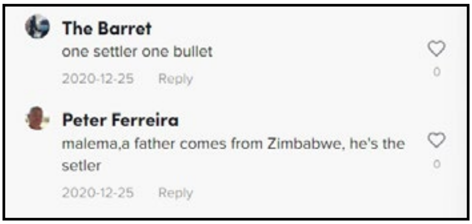
6

Like · Reply · 1y

↳ 5 Replies

Figure 2.1.14

We see further signs of the potency of xenophobia in the frequency with which claims of foreignness are used to silence. When a participant in a discussion on Facebook came out in support of the EFF protests in Brackenfell, he was immediately told “you’re a Zimbabwean not even Sout African so shut up“. Another user jumped in: “Well that fits […] All the thugs helping JUJU [a reference to Julius Malema] come from Zim because they cheap enough to afford and gullible when desperate“. Here and in 2.1.15 we see an effort to intimidate and silence an individual, as well as to taint the EFF with foreignness. We see a similar process in Figure 2.1.16 and the response in Figure 2.1.17, as well as in Figure 2.1.18. All of these examples point to the polarised nature of discussion on these threads, but also the way in which the discourse of the settler – used repeatedly by the EFF – is turned against it. Even if these claims are spurious, they have been taken up with vigour.



The Barret
one settler one bullet
2020-12-25 Reply
0

Peter Ferreira
malema,a father comes from Zimbabwe, he's the setler
2020-12-25 Reply
0

Figure 2.1.15

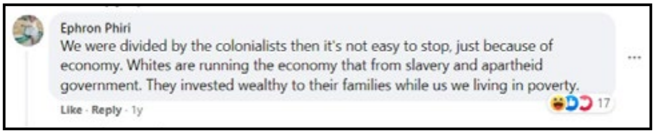

Ephron Phiri
We were divided by the colonialists then it's not easy to stop, just because of economy. Whites are running the economy that from slavery and apartheid government. They invested wealthy to their families while us we living in poverty.
Like · Reply · 1y 17

Figure 2.1.16

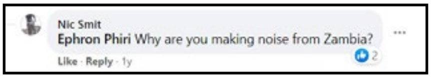
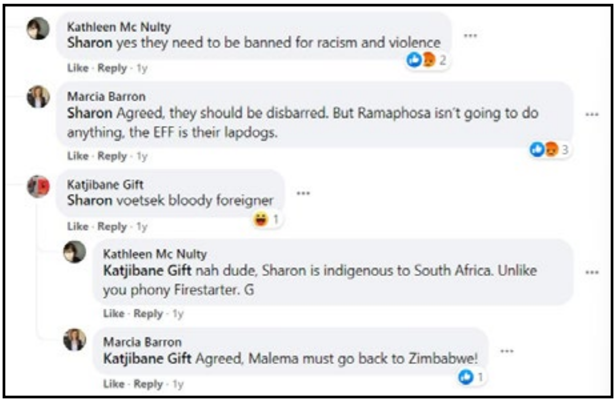

Nic Smit
Ephron Phiri Why are you making noise from Zambia?
Like · Reply · 1y 2

Figure 2.1.17


Kathleen Mc Nulty
Sharon yes they need to be banned for racism and violence
Like · Reply · 1y 2

Marcia Barron
Sharon Agreed, they should be disbarred. But Ramaphosa isn't going to do anything, the EFF is their lapdogs.
Like · Reply · 1y 3

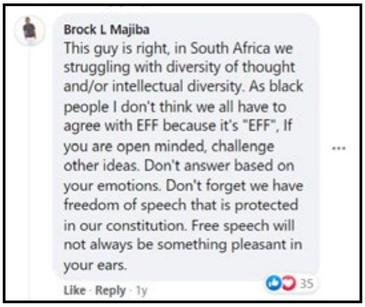
Katjibane Gift
Sharon voetsek bloody foreigner
Like · Reply · 1y 1

Kathleen Mc Nulty
Katjibane Gift nah dude, Sharon is indigenous to South Africa. Unlike you phony Firestarter. G
Like · Reply · 1y

Marcia Barron
Katjibane Gift Agreed, Malema must go back to Zimbabwe!
Like · Reply · 1y 1

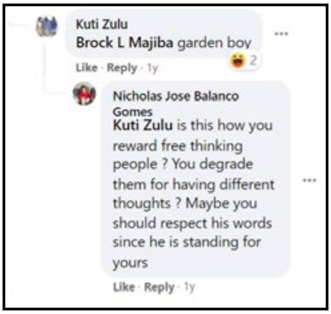
Figure 2.1.18

Interestingly, the discourse of foreignness is not limited to efforts to silence EFF supporters and vice versa. Figures 2.1.19 to 2.1.21 are extracts from a discussion found on a Facebook page of eNCA news. The page contains a video clip by the leader of the Cape Party in which he claims that the EFF is racist. An ostensibly black user responded with the post seen in Figure 2.1.19. This elicited a racist epithet from another user (Figure 2.1.20) and then another claim of foreignness (Figure 2.1.21).



Brock L Majiba
This guy is right, in South Africa we struggling with diversity of thought and/or intellectual diversity. As black people I don't think we all have to agree with EFF because it's "EFF", If you are open minded, challenge other ideas. Don't answer based on your emotions. Don't forget we have freedom of speech that is protected in our constitution. Free speech will not always be something pleasant in your ears.
Like · Reply · 1y 35

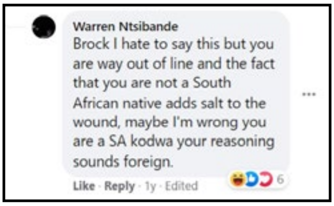
Figure 2.1.19



Kuti Zulu
Brock L Majiba garden boy
Like · Reply · 1y 2

Nicholas Jose Balanco Gomes
Kuti Zulu is this how you reward free thinking people ? You degrade them for having different thoughts ? Maybe you should respect his words since he is standing for yours
Like · Reply · 1y

Figure 2.1.20



Warren Ntsibande
Brock I hate to say this but you are way out of line and the fact that you are not a South African native adds salt to the wound, maybe I'm wrong you are a SA kodwa your reasoning sounds foreign.
Like · Reply · 1y · Edited 6

Figure 2.1.21

While much of the above may seem like a digression into a discussion of silencing, almost every one of these efforts to silence instead led to further engagement with these posts. This paradoxically drives the production of new user-generated content and increases time spent on the platform. Likewise, the act of trying to silence the EFF on social media had the contradictory effect of simply expanding the EFF's social media footprint during the Senekal and Brackenfell protests. The vitriolic nature of the discussion drew a great deal of additional traffic.

The EFF's substantial presence on Twitter is not remarkable considering how the party has always positioned itself when it comes to racism. Since its inception in 2013, the EFF has focused on land expropriation and economic emancipation, with race and class as central features. A good example of this is when the EFF led a demonstration to the Johannesburg Stock Exchange (JSE) in October 2015 as part of its 2016 municipal election campaign. The march was organised around the issue of economic emancipation but the narrative that emerged once the demonstration was over was that the JSE was racist and had treated those who had participated in the peaceful demonstration like criminals because they were black.[65] Over the years, the EFF has given the issue of racism more prominence, with the party being vocal or organising marches against many race-related events, including the Penny Sparrow racism case. It is this foregrounding of race issues that has partly facilitated their dominance on Twitter because, as explained by Bonilla and Rosa:

> ***Twitter affords a unique platform for collectively identifying, articulating, and contesting racial injustices from the in-group perspectives of racialized populations. Whereas in most mainstream media contexts the experiences of racialized populations are overdetermined, stereotyped, or tokenized, social-media platforms such as Twitter offer sites for collectively constructing counter narratives and re-imagining group identities.***[66]

As an opposition party that defines itself as a "radical and militant economic emancipation movement",[67] fighting against the exploitation of the black working class by the elite, the EFF is viewed as part of a counter-public that contests mainstream narratives, both politically and in the media. Although such counter publics have a long history in South Africa and elsewhere, the use of social-media platforms such as Twitter by young black South Africans has shown the potential of such spaces to organise, galvanise debates, and influence public action around injustices such as racism.[68] Notably, Suzanne Beukes also contends that Twitter needs to be viewed as a consensus-building medium because it offers people the opportunity to assemble around a particular knowledge base that comes pre-packaged with certain arguments.[69] It was through consensus-building that the student movements of 2015/2016, organised around the hashtags #RhodesMustFall and #FeesMustFall, were able to come together to constitute and link to a particular digital public.[70] For oppositional parties like the EFF, consensus

[65] Greg Nicolson, 'EFF marches: 'This isn't a Mickey Mouse organisation", *Daily Maverick,* 28 October 2015. https://www.dailymaverick.co.za/article/2015-10-28-eff-marches-this-isnt-a-mickey-mouse-organisation/

[66] Yarimar Bonilla and Jonathan Rosa, "# Ferguson: Digital protest, hashtag ethnography, and the racial politics of social media in the United States, in *American ethnologist*, Vol. 42, No. 1 (2015), pp. 4–17, at page 7.

[67] https://www.politicsweb.co.za/news-and-analysis/founding-manifesto-of-the-economic-freedom-fighter

[68] Suzanne Beukes, 'An Exploration of the Role of Twitter in the Discourse Around Race in South Africa. Using the# Feesmustfall Movement as a Pivot for Discussion' in Urte Undine Frömming , Steffen Köhn, Samantha Fox and Mike Tery (eds.), *Digital Environments. Ethnographic Perspectives across Global Online and Offline Spaces.* (Bielefeld: transcript, 2017), pp. 195–210, at page 195.

[69] Ibid., at page 205.

[70] Tanja Bosch, 'Twitter and participatory citizenship:# FeesMustFall in South Africa' in Bruce Mutsvairo (ed.), *Digital activism in the social media era* (Cham: Palgrave Macmillan, 2016), pp. 159–173, at page 164.

building becomes a useful tool for mobilising support on Twitter as these social-media platforms are designed to encourage "broadcasting over engagements, posts over discussions, shallow comments over deep conversations" and this is effective for political mobilisation but can also engender polarisation.[71]

Unsurprisingly then, given the central role played by the EFF in the protests at Senekal and Brackenfell and the dynamics described above, the EFF was the key driver of social media traffic on Twitter relating to these two events. The party sat at the intersection of posts relating to the two events (see Figure 2.1.22). Our research suggests that this is also true across Facebook and TikTok in relation to these flashpoints. The EFF's dominance of traffic was only challenged by Julius Malema's own personal Twitter account; both generated much more traffic than media outlets such as Radio 702 and eNCA and other political parties like the Democratic Alliance.

Despite the fact that these two flashpoints – Brackenfell and Senekal – focus on two ostensibly very different issues (racism in a school in the urban Western Cape; the killing of a farmer in the rural Free State), on social media the constant was a heightened presence of posts by, responses to, and posts about the EFF and their actions. When, for example, looking at the number of users in the Senekal dataset that have elicited responses, we can see the outsized influence of the EFF, followed once again by responses to Julius Malema's personal account (see Figure 2.1.23).

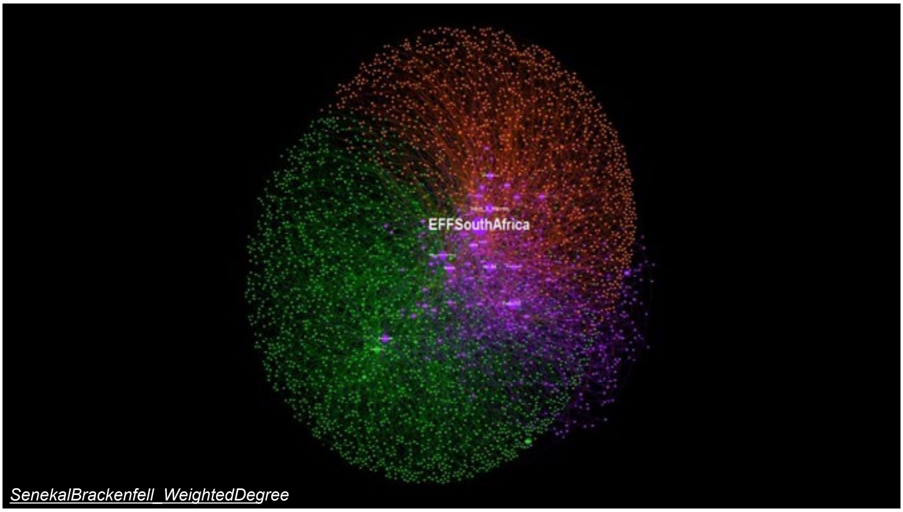


Figure 2.1.22: *Network analysis of datasets relating to the Senekal and Brackenfell protests. Senekal nodes are coloured green and Brackenfell nodes are coloured red. Nodes that are at the intersection of both datasets are in purple.*

[71] Suzanne Beukes, 'An Exploration of the Role of Twitter in the Discourse Around Race in South Africa. Using the# Feesmustfall Movement as a Pivot for Discussion' in Urte Undine Frömming , Steffen Köhn, Samantha Fox and Mike Tery (eds.), *Digital Environments. Ethnographic Perspectives across Global Online and Offline Spaces.* (Bielefeld: transcript, 2017), pp. 195–210, at page 206.

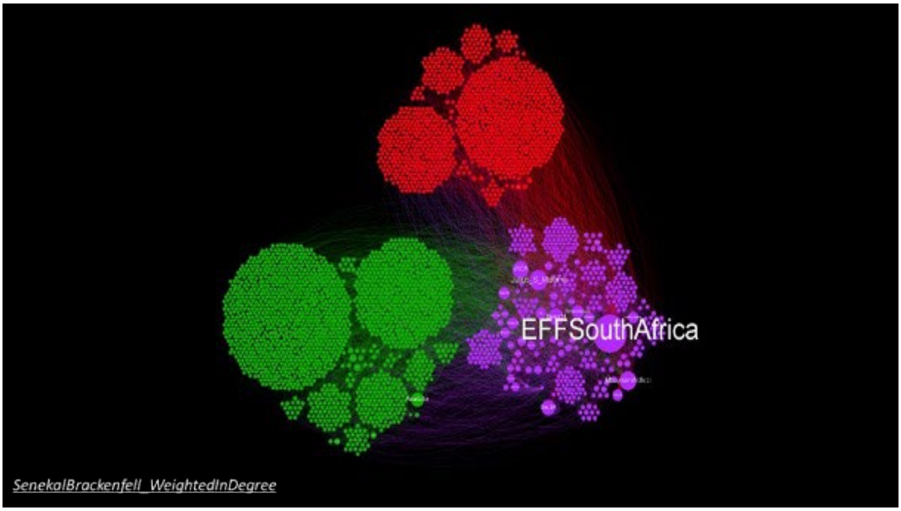


Figure 2.1.23: *Network analysis showing users being replied to in the Senekal and Brackenfell protest datasets.*

What also becomes clear when looking at Figure 2.1.23 is the very distinct groupings formed in these datasets. This echoes broader analyses of Twitter undertaken by Kyle Findlay who has highlighted the increasingly polarised nature of Twitter engagements in South Africa.[72] In an analysis based on 27 000 tweets about the Senekal protests (Figure 2.1.24), Findlay highlights what he refers to as ultra-polarised discussion where each group is in its own filter bubble and therefore dealing with a different set of facts and narratives. He notes that this is a key change in social media behaviour, as previously South African Twitter users would "all get our facts from the same place, even if our interpretations of those facts veer off in vastly different directions".[73]

[72] Findlay's review of Twitter usage in South Africa is tellingly titled "2020 in review: The year crude populism and polarisation took hold". https://www.superlinear.co.za/2020-in-review-the-year-crude-populism-and-polarisation-took-hold/

[73] Superlinear, '2020 in review: The year crude populism and polarisation took hold'. https://www.superlinear.co.za/2020-in-review-the-year-crude-populism-and-polarisation-took-hold/

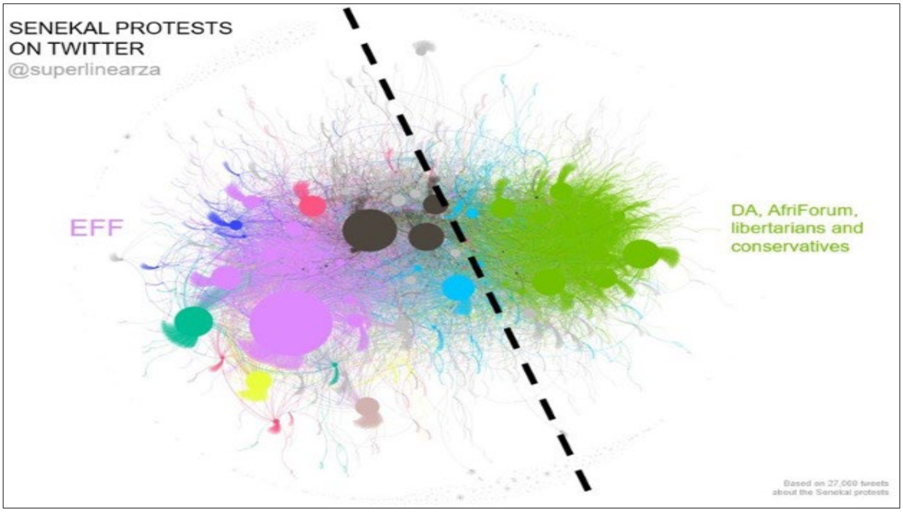


Figure 2.1.24.[74]

Similar features were evident in our qualitative analysis. Tellingly, our research shows that rather than being a place where users came to get news, the discussion generated by mainstream media content on social media instead resulted in the collision of very different 'facts' and narratives; these threads often rapidly became vitriolic in nature.

Findlay's analysis also shows that, despite forming a relatively small proportion of Twitter users, the white body politic and supporters of the EFF were by far the most vocal of social media users (see Figure 2.1.25). Given the eagerness of these two groups to spar on social media, it is clear that their dominance is in part due to their outsized engagement with the rival group. This once again highlights that, while the dynamics and rhetoric described in this report are concerning, it is important to remember that these are driven by a relatively small community of overall users. Yet at the same time these groups have played an outsize role in shifting our national narratives, because of their widening social media footprint.

[74] https://twitter.com/superlinearza/status/1326821843593007106

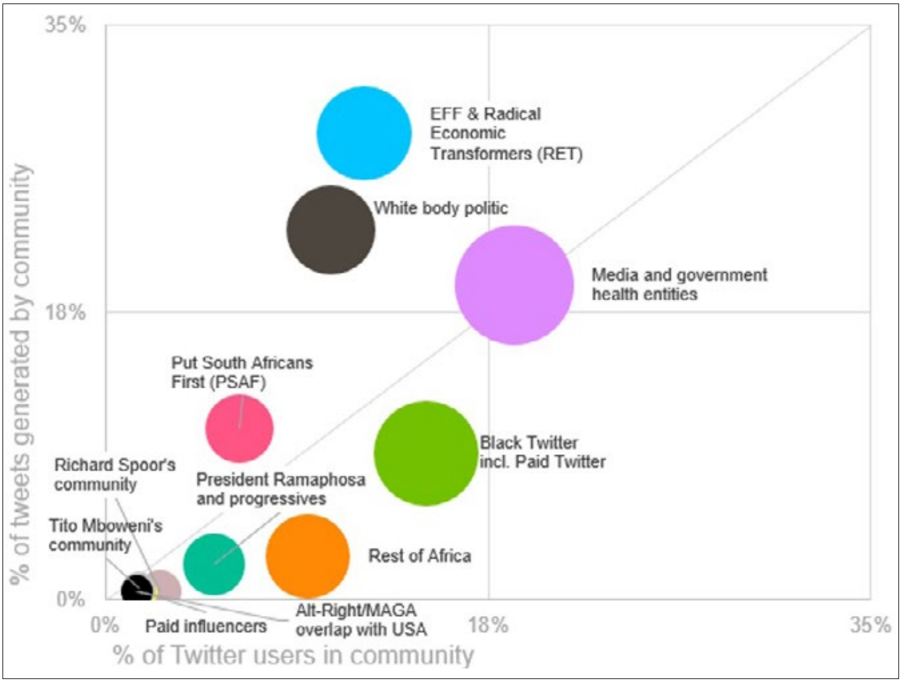


Figure 2.1.25: *Summary of the top communities discussing South African political and social issues. Communities above the diagonal line were the most vocal.*[75]

Findlay concludes that these trends point to "the hardening of race relations between far left and far right (read: black and white) South Africans".[76] This came through particularly strongly when someone was seen as crossing the ideological and racial divide (as seen in Figures 2.1.19 to 2.1.21). Such users became the targets of particularly vitriolic responses; they were depicted as 'race traitors'. We turn to this phenomenon in the next part of this study.

[75] Kyle Findlay, '2020 in review: The year crude populism and polarisation took hold'. https://www.superlinear.co.za/2020-in-review-the-year-crude-populism-and-polarisation-took-hold/

[76] Ibid.

2.2 The Race Traitor and the Sell-Out

The dominance of the 'white far right' and the 'black radical left' in much political discussion on social media has meant that race and political ideology are habitually conflated in online debate. This was visible at every turn in the flashpoints that we focused on. The assumption that white and black participants in these charged online discussions would cling to predictable and racially defined positions typically had consequences for those who did not do so. In the overheated language of social media, they were often branded as sell-outs and race traitors.

This vitriol was not reserved exclusively for those who crossed perceived racial and ideological lines. When users who were identified as Coloured or Indian entered discussions relating to the Brackenfell and Senekal protests, for example, they were immediately set upon by other users who indicated that they had no place in these debates and in South Africa more broadly. For example, on an eNews Channel Africa page on Facebook reporting on the Brackenfell protests, a user who was identified as Indian engaged in a heated back and forth with black interlocutors that quickly degenerated into claims and counterclaims about race. The former was told that Indians should not get in "the ring", and that his views were invalid (Figure 2.2.1). Similarly, an ostensibly Indian user who was critical of the EFF was told by a black user "shame I ddnt know the likes of you exist in such matters sitting on a fence camouflaging like chameleon waiting for whites to arrive turn white when Africans arrive you become an african get a grip!". The claim was made more pointedly with the statement that the "likes of you" do not exist in "such matters".

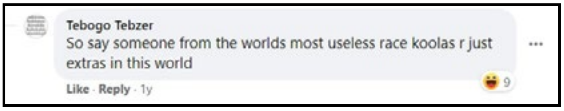
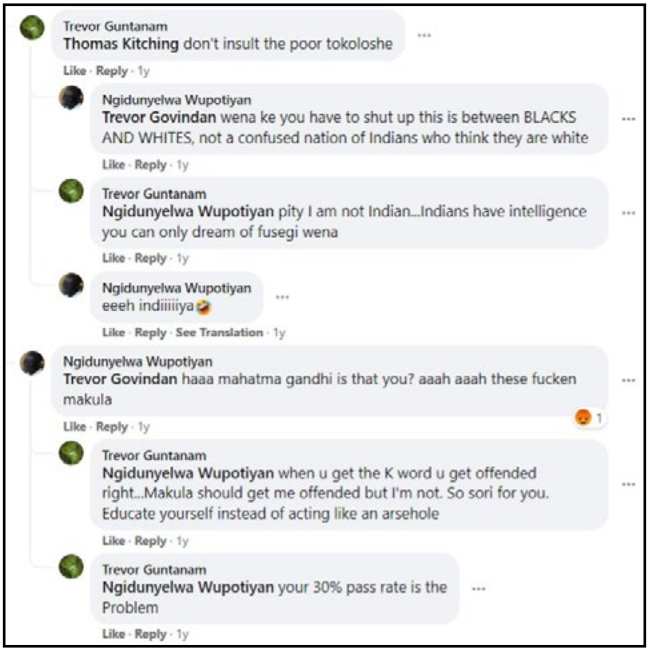

Tebogo Tebzer
So say someone from the worlds most useless race koolas r just extras in this world
Like · Reply · 1y
9

Figure 2.2.1

We see the same dynamic in Figure 2.2.2. The figure referred to as a "tokoloshe" here is Jack Markovitz, on whom more will be said below. While the descent of a conversation into slurs and insults is hardly surprising given the dynamics discussed in this report, the use of the term 'makula' shows the difficulties that the use of local terms can have in the interpretation of hate speech. Debate erupted around the meaning of the word in 2011 when Julius Malema, then still the leader of the ANC Youth League, used the word, often considered derogatory to Indians, when he addressed residents in the informal settlement of Thembelihle, situated in Lenasia. In his speech he stated "Bana ba lena ba tshwanetse ba dumelelwe gore ba tsene sekolo le bana ba makula mona [Your children must be allowed to go to school with Indian children]". Charges of hate speech were brought against him as the term is derived from the derogatory word 'coolie'. Many pointed out that this is often the only term used to describe Indian people in certain black communities, and that no offence is meant by the word. The term, in other words, can be innocuous or potent depending on context and inflexion.[77]



Trevor Guntanam
Thomas Kitching don't insult the poor tokoloshe
Like · Reply · 1y

Ngidunyelwa Wupotiyan
Trevor Govindan wena ke you have to shut up this is between BLACKS AND WHITES, not a confused nation of Indians who think they are white
Like · Reply · 1y

Trevor Guntanam
Ngidunyelwa Wupotiyan pity I am not Indian...Indians have intelligence you can only dream of fusegi wena
Like · Reply · 1y

Ngidunyelwa Wupotiyan
eeeh indiiiiiya🤣
Like · Reply · See Translation · 1y

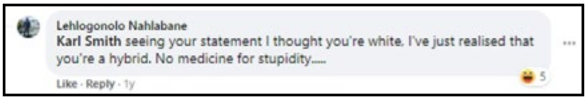
Ngidunyelwa Wupotiyan
Trevor Govindan haaa mahatma gandhi is that you? aaah aaah these fucken makula
😡 1
Like · Reply · 1y

Trevor Guntanam
Ngidunyelwa Wupotiyan when u get the K word u get offended right...Makula should get me offended but I'm not. So sori for you. Educate yourself instead of acting like an arsehole
Like · Reply · 1y

Trevor Guntanam
Ngidunyelwa Wupotiyan your 30% pass rate is the Problem
Like · Reply · 1y

Figure 2.2.2

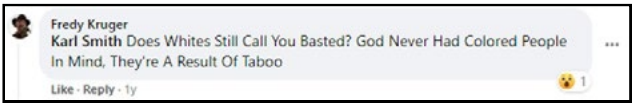
Those identified by other users as Coloured often faced identical treatment. In one striking instance, discussion of news reports about Brackenfell quickly transitioned into ad hominem attacks on a user that others assumed was Coloured. He was repeatedly instructed that Coloureds did not fit into the "correct" racial and social binaries (see Figures 2.2.3 and 2.2.4 for other related statements).



Lehlogonolo Nahlabane
Karl Smith seeing your statement I thought you're white. I've just realised that you're a hybrid. No medicine for stupidity.....
😆 5
Like · Reply · 1y

Figure 2.2.3

[77] See T Osiame Molefe, 'There is, thankfully, a Pedi word for big 'misunderstanding", *Daily Maverick*, 21 October 2011. https://www.dailymaverick.co.za/opinionista/2011-10-21-there-is-thankfully-a-pedi-word-for-big-misunderstanding/



Fredy Kruger
Karl Smith Does Whites Still Call You Basted? God Never Had Colored People In Mind, They're A Result Of Taboo

Figure 2.2.4

Those who did not fit into the binaries of white/right and black/left often encountered similar treatment. During the Brackenfell protests, Jack Markovitz was repeatedly identified as a "race traitor" on social media. This followed a widely distributed interview in which Markovitz, a white Jewish student at the University of Cape Town who is a member of the EFF, called the DA a white supremacist party, claimed Mandela "sold us out", and called for the transferral of "generational wealth and land to the disenfranchised people of apartheid". His comments elicited classic antisemitic responses of the kind seen in Figure 2.2.5, which impute an affinity between Jews and Communism.[78]

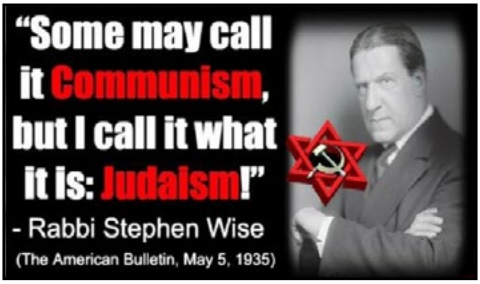

"Some may call it Communism, but I call it what it is: Judaism!"
- Rabbi Stephen Wise
(The American Bulletin, May 5, 1935)

Figure 2.2.5

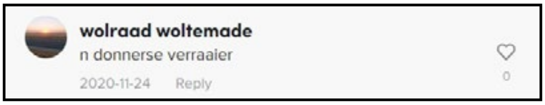
This was one of the few antisemitic posts that we found in our data relating to Markovitz, but reporting by various groups suggest that at the time his comments resulted in a deluge of antisemitic abuse on social media: "No surprise that he's a Jew, they've always supported terrorists"; "He is part of the new world order trying to control the world"; "Jewish youth voting for the SA Holocaust 2021"; and "Zionistiesejood [Zionist Jew], friend of Soros".[79] The absence of such material in our data from Twitter and Facebook points to the efficacy of content moderation that has retrospectively cleaned the online record (for how such content-moderation processes work, see Section Three).

What remained, however, were vicious attacks on Markovitz for being a "race traitor". For example, when a user named 'trotsesuidafrikaner' (proud South African) posted a clip on TikTok of the interview with the caption "What do you call

[78] Paul Boller, Jr. and John George, *They Never Said It: A Book of Fake Quotes, Misquotes, & Misleading Attributions* (New York & Oxford: Oxford University Press, 1989), page 132.

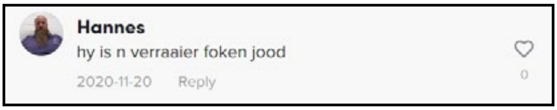
[79] Tali Feinberg, 'Antisemites and Jews see red over Markovitz's EFF comments', *Jewish Report*, 7 December 2020.

this guy?", the most common response was that he is a "verraaier [traitor]" (figures 2.2.6 to 2.2.9), with others suggesting a "wit kaffer [white kaffir]" and a "waist of white skin". Others wished death upon Markovitz.[80] The posts were no less vitriolic on Facebook (see, for example, Figure 2.2.10).[81]



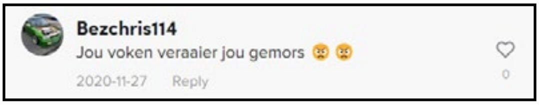
wolraad woltemade
n donnerse verraaier
2020-11-24 Reply
0

Figure 2.2.6



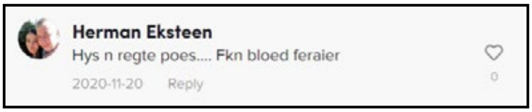
Hannes
hy is n verraaier foken jood
2020-11-20 Reply
0

Figure 2.2.7



Bezchris114
Jou voken veraaier jou gemors 😡 😡
2020-11-27 Reply
0

Figure 2.2.8



Herman Eksteen
Hys n regte poes.... Fkn bloed feraier
2020-11-20 Reply
0

Figure 2.2.9

[80] One user wrote, "Hopefully he gets the samething happen to him that happened to Amy biehl", in reference to the white anti-apartheid activist who was killed in 1993 in Gugulethu by black Africans while a crowd shouted anti-white slurs. Another wrote, "how sad his parents must be. kill it before it start breeding with another braindead eff shit turd". On another TikTok page, a screenshot of an image from a mobile phone is posted which suggests that he must be suffering from some form of foetal alcohol syndrome (see Figure 2.2.10).

[81] On a Taxi Times thread, one user wrotes "Shame! It must be hard to be so confused, to have such an identity uncertainty!" while numerous users referred to Markovitz as "Judas", with claims again appearing that he must have been paid. One user, for example, claimed "his getting paid 20K a month wot do we expect damn traitor". Another user wrote "Hier is vir jou a karretjie Hy ry lekker [here is a car for you, it rides well]" followed by a picture of a hearse. Whether this is a metaphor for Markovitz death as a white man for his supposedly traitorous actions or the wishing of his actual death is unclear.

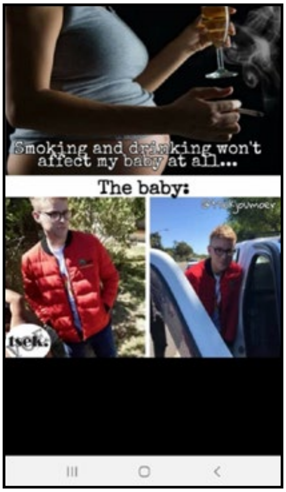


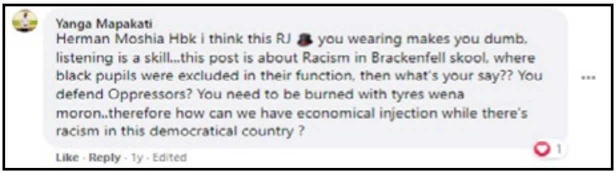
Figure 2.2.10

Markovitz became the lightning rod for the idea of the white race traitor, something that emerged in almost every thread where his name appeared. Yet white users were not alone in criticising Markovitz. Numerous black users on the same thread also described him as a traitor. They, however, identified the problem as Markowitz's membership of the EFF and the party's naivete for trusting white people.[82] Much of the above material was in response to posts from news organisations on Facebook. These Facebook posts often hosted the most heated and vitriolic engagements that we came across, many of which crossed the line into the realm of hate speech.

Similarly, when black users criticised the EFF, they were subject to derogatory remarks and racist stereotypes. On the same IOL news thread referred to above in which an Indian user was accused of "camouflaging like a chameleon", an ostensibly black user accused the EFF of being bullies. The response was immediate: he was a "white wannabe", "you're a disgraceful to human nature nxem", a "sell out", a "Yeeeeees Buss Boy [a yes baas boy]" and a "clever black who are actually morons by defending racism".

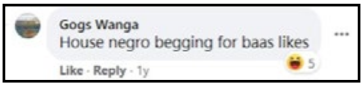
[82] One user flagged their distrust of Markovitz by stating "Steve Biko warned us about those that will come in our circles and behave as one of us whilst they not, they liberal in outlook…so beware!". Another user wrote "WHY EFF TRUST THIS GUY,WHO IS THIS IN THE FIRST PLACE,MXM THERE IS NO DIFFERENT FROM YOU AND ANC,ALWAYS TRUST THESE WHITE BOYS,WHAT MAKES MANDELA TO BECOME SOFT IS TO TRUST WHITE PEOPLE. MALEMA IS ABOUT TO DO THE SAME FUCKEN ERROR,SOBUKWE IS STILL MY FUCKEN HERO". Meanwhile a third user makes this distrust far more explicit, stating ""Oho…don't fall for this… he was sent by the same racist white people to pretend that he's against their sickness 😂😂😂. Oho we know these things…it doesn't mean if they walk with you and that they are for you. They are just confusing the enemy".

Here the term ‘clever blacks’ derives from President Jacob Zuma’s reprimand of Africans who are ‘too clever’ and ‘eloquent in criticising’ their own traditions. The term has since taken on a life of its own as a reference to the South African black middle class.[83] Once again, users employed derogatory language that would escape the attention of content moderators unfamiliar with the South African social and political landscape. Those who dared to question the wisdom of the protests at Brackenfell were heckled with the responses of the kind seen in figures 2.2.11 and 2.2.12. In these examples, we not only see the mobilisation of apartheid-era terms for collaborators but also the charge that those who do not support the policies of the black left are “not black enough”.



Yanga Mapakati
Herman Moshia Hbk i think this RJ 🎩 you wearing makes you dumb. listening is a skill...this post is about Racism in Brackenfell skool. where black pupils were excluded in their function. then what's your say?? You defend Oppressors? You need to be burned with tyres wena moron..therefore how can we have economical injection while there's racism in this democratical country ?

Figure 2.2.11

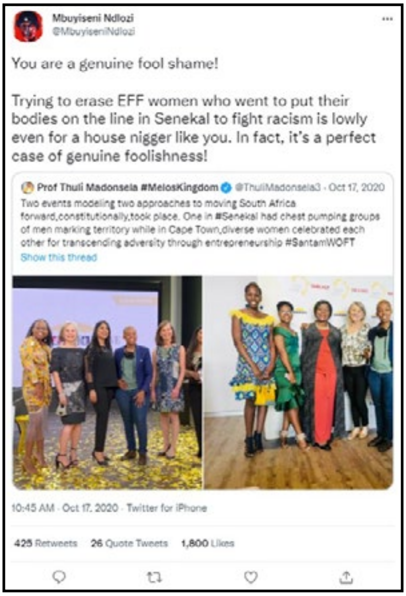


Gogs Wanga
House negro begging for baas likes

Figure 2.2.12

In much the same way that Markowitz became a favoured representative of the white ‘verraaier’, Thuli Madonsela (the former Public Protector) has become a favoured example of the ‘black sell-out’ on social media. For example, when Madonsela tweeted about the Santam Women of the Future awards – contrasting these to the “chest pumping groups of men marking territory“ in Senekal – this struck a nerve, as seen in the response by Mbuyeseni Ndlozi, an EFF Member of Parliament (Figure 2.2.13). Ndlozi repeated the use of the same derogatory term in other tweets which still remain on Twitter.

[83] E Dimitris Kitis, Tommaso M Milani and Erez Levon, “Black diamonds’, ‘clever blacks’ and other metaphors: Constructing the black middle class in contemporary South African print media’ in *Discourse & Communication*, Vol. 12, No. 2 (2018), pp. 149–170.



Mbuyiseni Ndlozi
@MbuyiseniNdlozi

You are a genuine fool shame!

Trying to erase EFF women who went to put their bodies on the line in Senekal to fight racism is lowly even for a house nigger like you. In fact, it's a perfect case of genuine foolishness!

Prof Thuli Madonsela #MelosKingdom @ThuliMadonsela3 · Oct 17, 2020
Two events modeling two approaches to moving South Africa forward,constitutionally,took place. One in #Senekal had chest pumping groups of men marking territory while in Cape Town,diverse women celebrated each other for transcending adversity through entrepreneurship #SantamWOFT
Show this thread

10:45 AM · Oct 17, 2020 · Twitter for iPhone

425 Retweets 26 Quote Tweets 1,800 Likes

Figure 2.2.13

The theme was immediately taken up by other black users. One claimed that she was using her position as a "stepping stone to whiteness"; another questioned her blackness ("She is a black white woman,you can't be sleeping with the devil and be a saint, she is one of them, trapped in the black skin, maybe she has started bleaching also so she can be pink enough"). These derogatory comments were not limited to Twitter, and appear with a disturbing frequency: "she is a white people agent", a "white puppet",[84] she had been "Captured by WMC [White Monopoly Capital]", she was an apartheid apologist, a claim made with dramatic visual effect by the circulation of a fake image of Madonsela stood next to the last apartheid President FW De Klerk in a dress representing the apartheid South African flag (see Figure 2.2.14).[85]

In a social media environment dominated by a white far right that both opposes and draws its strength from a black radical left (and vice versa), those who do not fit into these neat categories are favoured targets. By assailing them, the white far right and black radical left not only drive traffic but seek to undermine those who do not hew to their worldviews, as well as to demarcate and police the boundaries of their rival camps.

[84] Mpho Sibanyoni, 'Meet the protector of Thuli Madonsela's heart', *SowetanLIVE*, 27 July 2018. https://www.sowetanlive.co.za/news/south-africa/2018-07-27-meet-the-protector-of-thuli-madonselas-heart/

[85] https://factcheck.afp.com/no-thuli-madonsela-did-not-pose-apartheid-era-flag-fw-de-klerk-its-forgery.

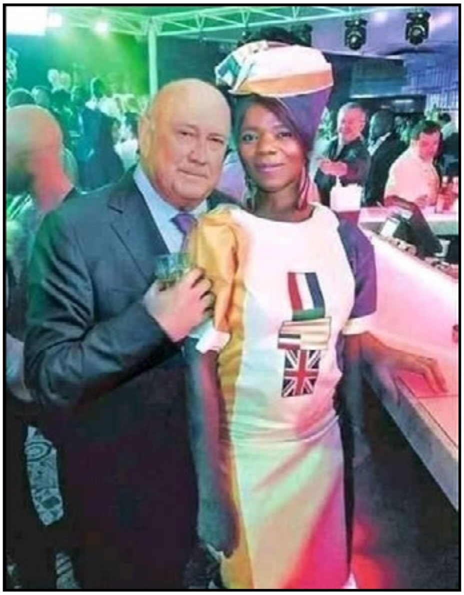
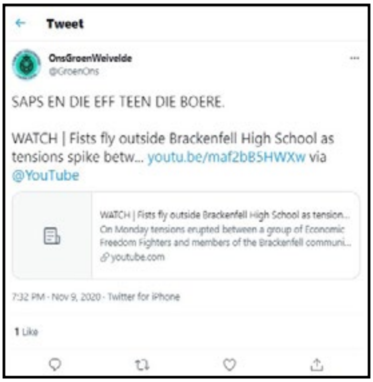


Figure 2.2.14

2.3 White Victimhood: Stories of Decay, Degeneration, and Peril

Political discussion on social media in South Africa often takes on an Alice in Wonderland quality where real-world economic and social realities appear in distorted and fantastical forms. Representative of this is the recurrent motif on Twitter and Facebook of white South Africans as an imperilled group, the primary victims of misrule, decay, and decline in South Africa. In its more extreme manifestation, this takes on the form of narratives of 'white genocide'. This white victim complex, propagated by the white far right, typically draws on nostalgia for the past and portrays black South Africans as criminogenic and incapable of good governance. White South Africans, by contrast, are positioned as a group without which the South African economy will crumble and fall. Unsurprisingly, these ideas were a constant and prominent presence in content relating to the Brackenfell and Senekal protests.

Typical of the claim that whites are oppressed in the new South Africa is a tweet (Figure 2.3.1) asserting that the police and EFF were "against the Boers" during the Brackenfell protests. This claim was repeated on numerous occasions on the Facebook group of #BlackMonday, a movement designed to bring awareness to farm murders. The profile image of the group can be seen in Figure 2.3.2: stylised white crosses in a rural landscape conveying a claim of rootedness in the land and signifying farmers who have suffered violent deaths. The use of the bloody handprint – as seen in the BlackMonday logo – is a constant theme across various farm murder groups.



Tweet

OnsGroenWelvelde
@GroenOns

SAPS EN DIE EFF TEEN DIE BOERE.

WATCH | Fists fly outside Brackenfell High School as tensions spike betw... youtu.be/maf2bB5HWXw via @YouTube

WATCH | Fists fly outside Brackenfell High School as tension...
On Monday tensions erupted between a group of Economic Freedom Fighters and members of the Brackenfell communi...
youtube.com

7:32 PM · Nov 9, 2020 · Twitter for iPhone

1 Like

Figure 2.3.1

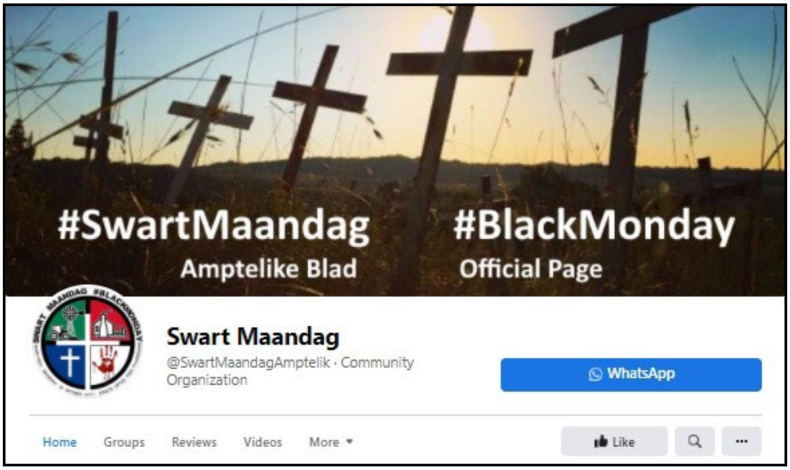


Figure 2.3.2

#BlackMonday was one of several ‘farm murder’ groups that proliferated online following the Senekal protests, such as ‘Farmers’ Lives Matter SA’ (almost 100 000 users) and the ‘Stop Farm Murders Movement’ (over 80 000 members). These particular groups (whose Facebook profile pages can be seen in Figures 2.3.3 and 2.3.4) positioned themselves in very different ways. Yet the content within these groups often follows very similar patterns: graphic imagery depicting white victims of violence, claims that farm murders are being ignored or denied, and calls for retribution against perpetrators.

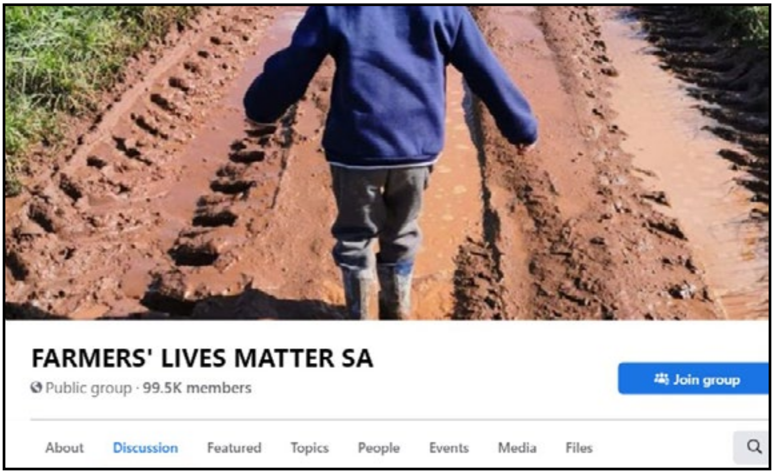
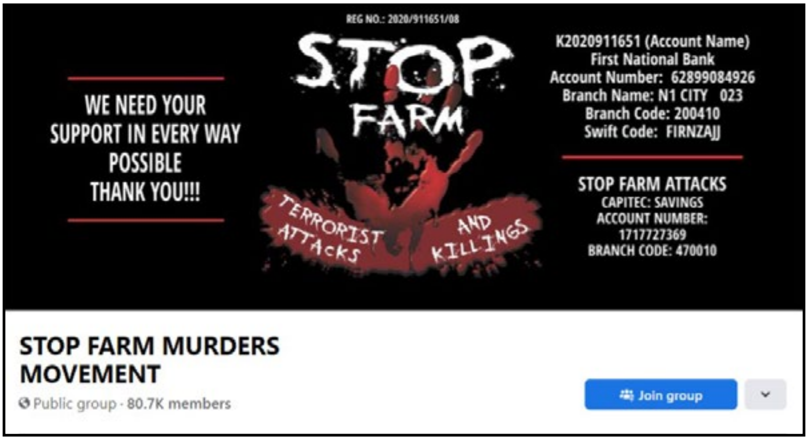


Figure 2.3.3



REG NO.: 2020/911651/08

STOP FARM TERRORIST ATTACKS AND KILLINGS

WE NEED YOUR
SUPPORT IN EVERY WAY
POSSIBLE
THANK YOU!!!

K2020911651 (Account Name)
First National Bank
Account Number: 62899084926
Branch Name: N1 CITY 023
Branch Code: 200410
Swift Code: FIRNZAJJ

STOP FARM ATTACKS
CAPITEC: SAVINGS
ACCOUNT NUMBER:
1717727369
BRANCH CODE: 470010

STOP FARM MURDERS
MOVEMENT

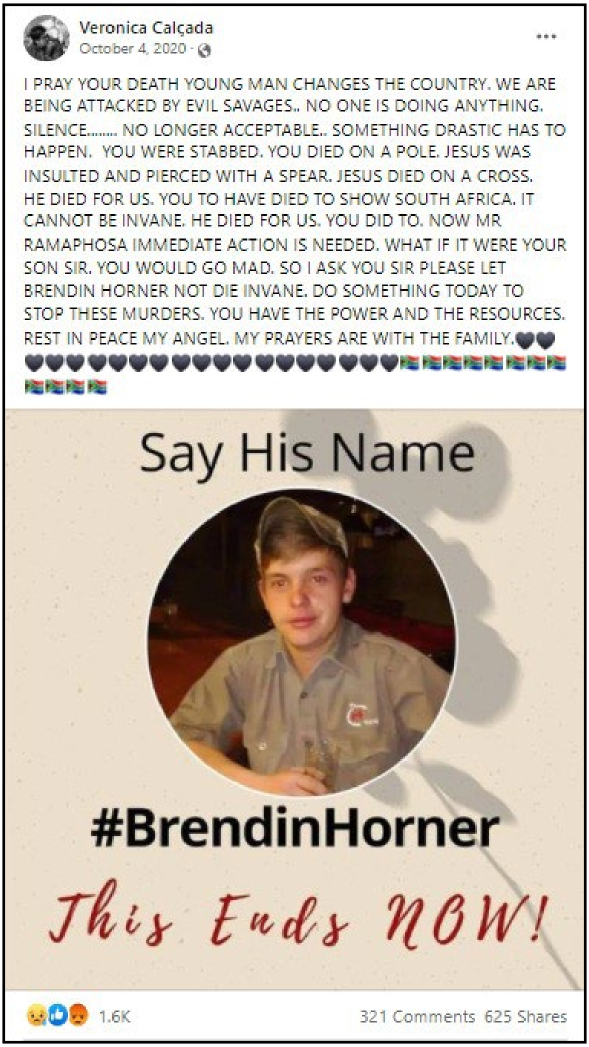
Public group · 80.7K members

Figure 2.3.4

The latter took on an extreme form following the arrest of those accused of murdering Brendin Horner: “Death Penalty - Klaar - Torture them and hang them slowly - Eye for an Eye“, “Hulle moet gemartel word tot hulle vrek die gemors [this rubbish should be tortured until they die]“, “Behead them and their Relative’s IN FULL PUBLIC VIEW ,, No BAIL , AND NO JAIL !!!!“ These sentiments are indicative of the strength of feeling generated by this particular event, but also of the dynamics which are possible when like-minded individuals band together on social media. Such baying for blood suggests the ways that Facebook groups can foment extreme responses, as users, unconstrained by the guardrails that typically contain discourse in the real world, egg each other on and adopt overheated language. Empathy is often an early victim in the frenzy.

These and other posts on these Facebook pages reflect an absence of faith in the government to mete out justice, as well as an assumption that those who bear the brunt of the failure of the government to enforce the law are white South Africans. These two ideas were also present on a range of other pages without obvious links to farm murders and white genocide. For example, a group called #OurVoices, which had 259 500 members and describes itself as “a caring, loving and non violent movement“ also became the site of emotive and racialised pleas for justice, railing against the government, and a call to arms after Horner’s murder (see, for example, figures 2.3.5 and 2.3.6).[86] Numerous posts referred to “evil barbarians“, “savages“, and “brutal thugs“ to the point that these descriptions became commonplace. Others blamed the rise of a black-led democratic government for the fall of South Africa (see, for example, figures 2.3.7 and 2.3.8).

[86] https://www.facebook.com/groups/ShutSADown/about

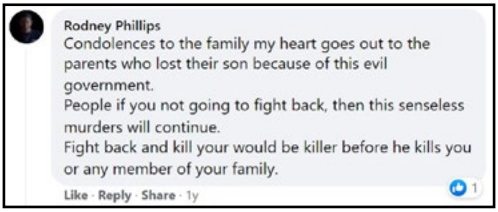


Veronica Calçada
October 4, 2020 ·

I PRAY YOUR DEATH YOUNG MAN CHANGES THE COUNTRY. WE ARE BEING ATTACKED BY EVIL SAVAGES.. NO ONE IS DOING ANYTHING. SILENCE........ NO LONGER ACCEPTABLE.. SOMETHING DRASTIC HAS TO HAPPEN. YOU WERE STABBED. YOU DIED ON A POLE. JESUS WAS INSULTED AND PIERCED WITH A SPEAR. JESUS DIED ON A CROSS. HE DIED FOR US. YOU TO HAVE DIED TO SHOW SOUTH AFRICA. IT CANNOT BE INVANE. HE DIED FOR US. YOU DID TO. NOW MR RAMAPHOSA IMMEDIATE ACTION IS NEEDED. WHAT IF IT WERE YOUR SON SIR. YOU WOULD GO MAD. SO I ASK YOU SIR PLEASE LET BRENDIN HORNER NOT DIE INVANE. DO SOMETHING TODAY TO STOP THESE MURDERS. YOU HAVE THE POWER AND THE RESOURCES. REST IN PEACE MY ANGEL. MY PRAYERS ARE WITH THE FAMILY.🖤🖤🖤🖤🖤🖤🖤🖤🖤🖤🖤🖤🖤🖤🖤🖤🖤🖤🖤🖤🇿🇦🇿🇦🇿🇦🇿🇦🇿🇦🇿🇦🇿🇦🇿🇦🇿🇦🇿🇦🇿🇦🇿🇦

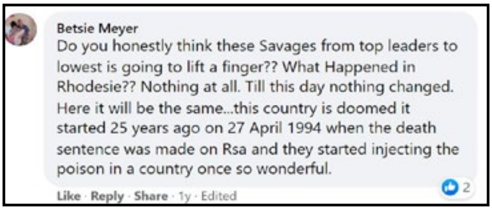
1.6K 321 Comments 625 Shares

Figure 2.3.5



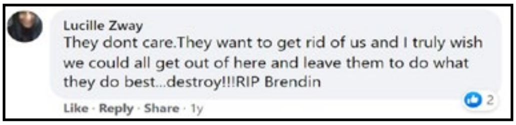
Rodney Phillips
Condolences to the family my heart goes out to the parents who lost their son because of this evil government.
People if you not going to fight back, then this senseless murders will continue.
Fight back and kill your would be killer before he kills you or any member of your family.

Figure 2.3.6



Betsie Meyer
Do you honestly think these Savages from top leaders to lowest is going to lift a finger?? What Happened in Rhodesie?? Nothing at all. Till this day nothing changed. Here it will be the same...this country is doomed it started 25 years ago on 27 April 1994 when the death sentence was made on Rsa and they started injecting the poison in a country once so wonderful.

Figure 2.3.7



Lucille Zway
They dont care.They want to get rid of us and I truly wish we could all get out of here and leave them to do what they do best...destroy!!!RIP Brendin

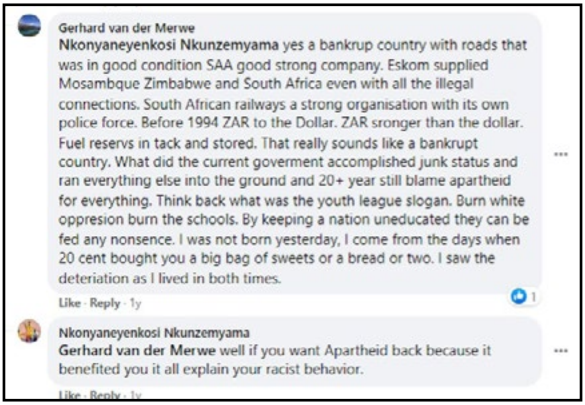
Figure 2.3.8

What is notable throughout these posts is that the 'they' and the 'we' that they refer to are rarely made explicit. This is a common feature on these sites. By consistently referring to blacks as 'they' and the other side of the dichotomy as 'we' the supposedly embattled white population is presented as victims of the new order. Equally importantly, these locutions save users from being flagged by content moderators who do not see obvious evidence of racism on these pages.[87]

Common too was the claim that South Africa has begun an inexorable decline, with this juxtaposed against the idea that South Africa was once wonderful. This takes a variety of forms ranging from nostalgia for past security and stability to a more extreme longing for the "good days" of apartheid. This latter view is encapsulated in a Facebook post in Figure 2.3.9 extracted from a page in the Farmer's Weekly SA Facebook group. Here the user systematically goes through various sectors of the South African economy under apartheid, suggesting that it was a well-oiled machine that has now been run into the ground by a government that erroneously blames apartheid for this state of affairs. This constant reference to a better past

[87] Jessica Ann Barraclough, Facebook's 'White Genocide' Problem: A Sociotechnical Exploration of Problematic Information, Shareability, and Social Correction in a South African Context (Masters Dissertation, University of Cape Town, 2020), at page 101.

(and the rapid disavowal that this was the case by black users) was a recurrent feature of the content we examined.

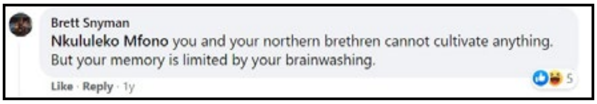


Gerhard van der Merwe
Nkonyaneyenkosi Nkunzemyama yes a bankrup country with roads that was in good condition SAA good strong company. Eskom supplied Mosambque Zimbabwe and South Africa even with all the illegal connections. South African railways a strong organisation with its own police force. Before 1994 ZAR to the Dollar. ZAR sronger than the dollar. Fuel reservs in tack and stored. That really sounds like a bankrupt country. What did the current goverment accomplished junk status and ran everything else into the ground and 20+ year still blame apartheid for everything. Think back what was the youth league slogan. Burn white oppresion burn the schools. By keeping a nation uneducated they can be fed any nonsence. I was not born yesterday, I come from the days when 20 cent bought you a big bag of sweets or a bread or two. I saw the deteriation as I lived in both times.

Nkonyaneyenkosi Nkunzemyama
Gerhard van der Merwe well if you want Apartheid back because it benefited you it all explain your racist behavior.

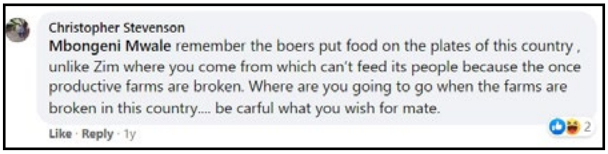
Figure 2.3.9

This narrative often manifested through the invocation of Zimbabwe as a warning, and the suggestion that black Africans are incapable of running South Africa. See, Figures 2.3.10 and 2.3.11, for example, which appeared on a *Daily Maverick* Facebook page reporting on the Senekal protests.



Brett Snyman
Nkululeko Mfono you and your northern brethren cannot cultivate anything. But your memory is limited by your brainwashing.

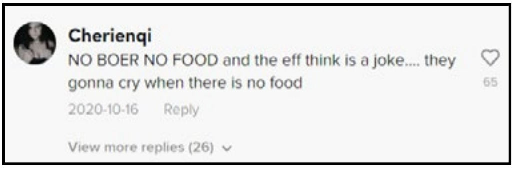
Figure 2.3.10



Christopher Stevenson
Mbongeni Mwale remember the boers put food on the plates of this country , unlike Zim where you come from which can't feed its people because the once productive farms are broken. Where are you going to go when the farms are broken in this country.... be carful what you wish for mate.

Figure 2.3.11

The claims that Africans cannot govern and that South Africa's survival is dependent on whites is made even more explicitly in a thread on TikTok in response to a short clip of EFF protesters in Senekal (see figures 2.3.12 to 2.3.15). Similarly, the productiveness of white South Africans was constantly highlighted by users who claimed that whites achieved their position in society solely because of hard work.[88]



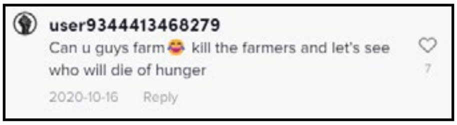
Cherienqi

NO BOER NO FOOD and the eff think is a joke.... they gonna cry when there is no food

2020-10-16 Reply

View more replies (26)

Figure 2.3.12

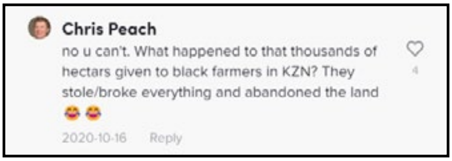


user9344413468279

Can u guys farm😂 kill the farmers and let's see who will die of hunger

2020-10-16 Reply

Figure 2.3.13

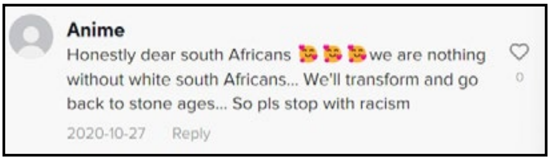


Chris Peach

no u can't. What happened to that thousands of hectars given to black farmers in KZN? They stole/broke everything and abandoned the land 😂😂

2020-10-16 Reply

Figure 2.3.14



Anime

Honestly dear south Africans 🥰🥰🥰we are nothing without white south Africans... We'll transform and go back to stone ages... So pls stop with racism

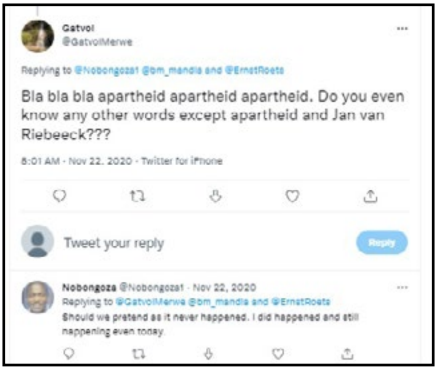
2020-10-27 Reply

Figure 2.3.15

Though some social media users trafficked only in implicit racism, others were less reticent. Some of this material has escaped content moderation. A user who described himself as a resident of Brackenfell, for example, engaged in a coarse three-day-long racist rant that still appears online. When challenged, he responded with the tweet seen in Figures 2.3.16 and 2.3.17. Here we see the familiar claim that a black government is the cause of South Africa's ills; they "fucked up this country IN ONLY 25 years". In posts like these, the cause for South Africa's problems

[88] Yves Vanderhaeghen, *Afrikaner Identity: Dysfunction and Grief* (Pietermaritzburg: University of KwaZulu-Natal Press, 2018), at pages 194–195.

becomes black Africans, a rhetorical strategy that denies the role of apartheid in South Africa’s present-day problems.



Gatvol
@GatvolMerwe

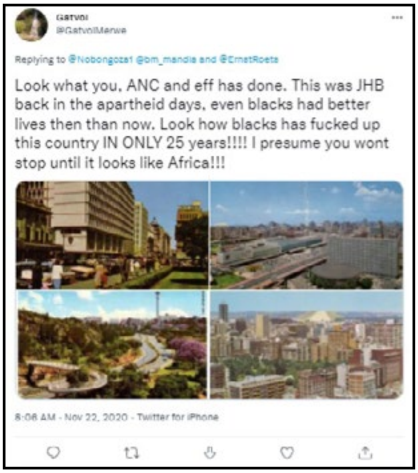
Replying to @Nobongoza1 @bm_mandla and @ErnstRoets

Bla bla bla apartheid apartheid apartheid. Do you even know any other words except apartheid and Jan van Riebeeck???

8:01 AM · Nov 22. 2020 · Twitter for iPhone

Nobongoza @Nobongoza1 · Nov 22, 2020
Replying to @GatvolMerwe @bm_mandla and @ErnstRoets
Should we pretend as it never happened. I did happened and still happening even today.

Figure 2.3.16



Gatvol
@GatvolMerwe

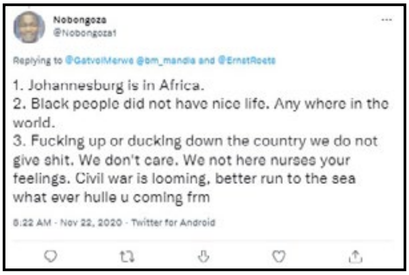
Replying to @Nobongoza1 @bm_mandla and @ErnstRoets

Look what you. ANC and eff has done. This was JHB back in the apartheid days. even blacks had better lives then than now. Look how blacks has fucked up this country IN ONLY 25 years!!!! I presume you wont stop until it looks like Africa!!!

8:06 AM · Nov 22, 2020 · Twitter for iPhone

Figure 2.3.17

We also see in this interaction the ways in which this nostalgic vision of the past and the refusal to acknowledge the lasting harms of apartheid can lead to a rapid escalation in online engagements. Figure 2.3.17 generated a furious response (Figure 2.3.18): “Civil war is looming, better run to the sea“, and a rare moment of agreement (Figure 2.3.19). Within a few short posts, both users draw lines in the sand and can only agree on the inevitability of civil war.



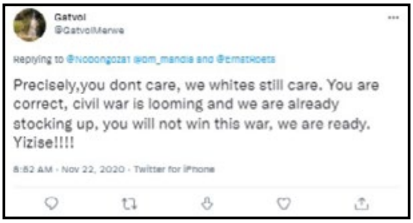
Nobongoza
@Nobongoza1

Replying to @GatvolMerwe @bm_mandla and @ErnstRoets

1. Johannesburg is in Africa.
2. Black people did not have nice life. Any where in the world.
3. Fucking up or ducking down the country we do not give shit. We don't care. We not here nurses your feelings. Civil war is looming, better run to the sea what ever hulle u coming frm

6:22 AM · Nov 22, 2020 · Twitter for Android

Figure 2.3.18



Gatvol
@GatvolMerwe

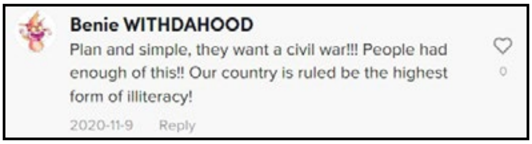
Replying to @Nobongoza1 @bm_mandla and @ErnstRoets

Precisely,you dont care, we whites still care. You are correct, civil war is looming and we are already stocking up, you will not win this war, we are ready. Yizise!!!!

8:52 AM · Nov 22, 2020 · Twitter for iPhone

Figure 2.3.19

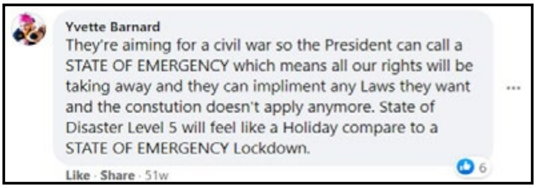
Talk of looming civil war and race war drew on a range of conspiratorial ideas that were stimulated by the protests at Senekal, fears of land expropriation, and Covid-19 lockdowns (see, for example, figures 2.3.20 to 2.3.23). While many feared the potential implications of a race war, others welcomed it (see, for example, Figure 2.3.24).



Benie WITHDAHOOD
Plan and simple, they want a civil war!!! People had enough of this!! Our country is ruled be the highest form of illiteracy!

2020-11-9 Reply

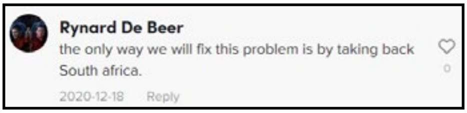
Figure 2.3.20



Yvette Barnard
They're aiming for a civil war so the President can call a STATE OF EMERGENCY which means all our rights will be taking away and they can impliment any Laws they want and the constution doesn't apply anymore. State of Disaster Level 5 will feel like a Holiday compare to a STATE OF EMERGENCY Lockdown.

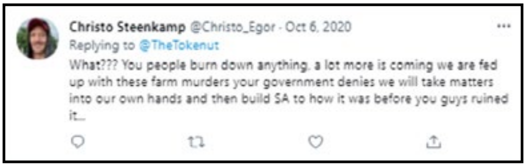
Like · Share · 51w

Figure 2.3.21



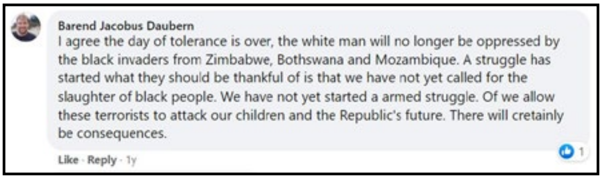
Rynard De Beer
the only way we will fix this problem is by taking back South africa.
2020-12-18 Reply

Figure 2.3.22



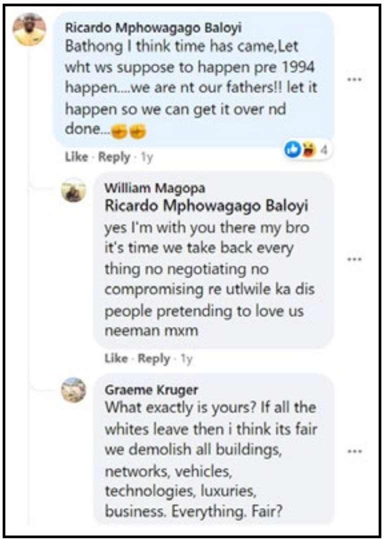
Christo Steenkamp @Christo_Egor · Oct 6, 2020
Replying to @TheTokenut
What??? You people burn down anything. a lot more is coming we are fed up with these farm murders your government denies we will take matters into our own hands and then build SA to how it was before you guys ruined it...

Figure 2.3.23



Barend Jacobus Daubern
I agree the day of tolerance is over, the white man will no longer be oppressed by the black invaders from Zimbabwe, Bothswana and Mozambique. A struggle has started what they should be thankful of is that we have not yet called for the slaughter of black people. We have not yet started a armed struggle. Of we allow these terrorists to attack our children and the Republic's future. There will cretainly be consequences.
Like · Reply · 1y 1

Figure 2.3.24



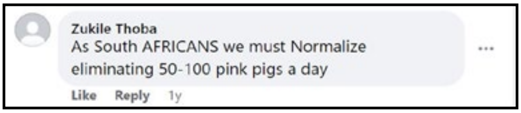
Ricardo Mphowagago Baloyi
Bathong I think time has came,Let wht ws suppose to happen pre 1994 happen....we are nt our fathers!! let it happen so we can get it over nd done...✊✊
Like · Reply · 1y 4

William Magopa
Ricardo Mphowagago Baloyi yes I'm with you there my bro it's time we take back every thing no negotiating no compromising re utlwile ka dis people pretending to love us neeman mxm
Like · Reply · 1y

Graeme Kruger
What exactly is yours? If all the whites leave then i think its fair we demolish all buildings, networks, vehicles, technologies, luxuries, business. Everything. Fair?

Figure 2.3.25

Worryingly, the expectation of the inevitability of a race war was also shared by some ostensibly black users (Figure 2.3.25). Though content openly calling for violence was the exception rather than the norm during the Senekal protests, Figure 2.3.26, a post from that period that is still present on the EFF's Facebook page, leaves little to the imagination. The fact that this post and others like it have not been removed by Facebook has only fuelled the idea within the far right that whites in South Africa are under siege and must 'fight back' to protect themselves from imminent genocidal violence.



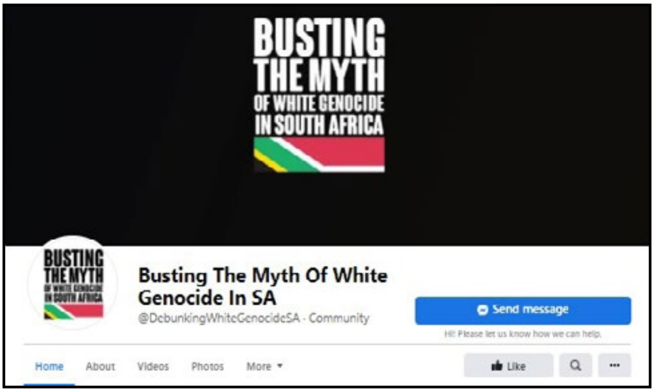
Zukile Thoba
As South AFRICANS we must Normalize eliminating 50-100 pink pigs a day

Figure 2.3.26

These fears about the vulnerability and victimhood of white South Africans, as well as the inevitability of race war, have drawn nourishment from the online activities of the EFF and its leader Julius Malema. Malema in particular has assumed a totemic place on social media. For white users on the far right, he has become the embodiment of both the 'swart gevaar' and the 'rooi gevaar' and the greatest threat to land, economy, and safety in South Africa.

2.4 Globalising Local Victimhood: Farm Murders

The targeting of white farmers has been taken up as a core issue by numerous white nationalist groups in South Africa and beyond to feed into, and act as evidence for, broader claims of a "white genocide". [89] These groups have found each other via, and collaborate on, social media. The mobilisation of the notion that whites in South Africa (and white Afrikaners in particular) are victims of a "white genocide" is hardly new but it has gained increasing traction in recent years and was a consistent feature of social-media content relating to the Senekal and Brackenfell protests. In fact, this notion of 'white genocide' is so widespread on social media that it has led to its debunking by both Africa Check and the BBC, as well as the creation of the "Busting the Myth of White Genocide in SA" Facebook page (figures 2.4.1 and 2.4.2).[90]



BUSTING THE MYTH OF WHITE GENOCIDE IN SOUTH AFRICA

Busting The Myth Of White Genocide In SA

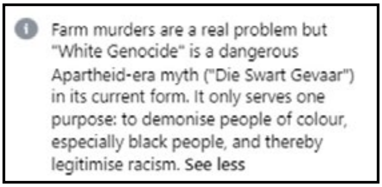
@DebunkingWhiteGenocideSA · Community

Hi! Please let us know how we can help.

Figure 2.4.1

[89] There are other groups in South Africa that are statistically at greater risk of murder than farmers, such as night-shift workers and Uber drivers. The number of farm murders has consistently been below one hundred per year (AfriForum reported 63 farm murders in 2020), with the national average of murders every day in South Africa numbering 58.4 per day between 1 April 2019 to 31 March 2020. AfriForum, 'AfriForum releases farm attack and murder statistics for 2020', 17 February 2021. https://afriforum.co.za/en/afriforum-releases-farm-attack-and-murder-statistics-for-2020/; Hannelie Marx Knoetze, 'Romanticising the "Boer": Narratives of White Victimhood in South African Popular Culture' in *Journal of Literary Studies,* Vol. 36, No. 4 (2020), pp. 48–69, at page 59.

[90] The group, which has just over 25 000 followers, is run by an anonymous group of administrators who describe themselves as activists from various racial and economic backgrounds. Their stated aim is to provide a platform for the sharing of credible and accurate facts and statistics regarding crime in South Africa and to "counter the onslaught of extreme Right-Wing propaganda and conspiracy theories flooding the internet and social media".

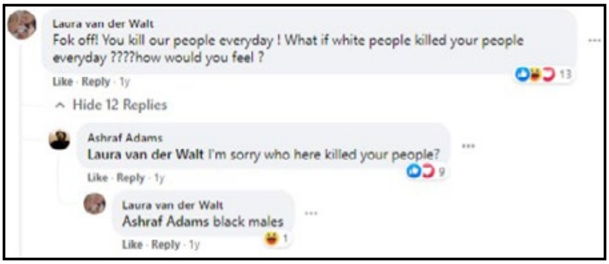


Farm murders are a real problem but "White Genocide" is a dangerous Apartheid-era myth ("Die Swart Gevaar") in its current form. It only serves one purpose: to demonise people of colour, especially black people, and thereby legitimise racism. See less

Figure 2.4.2

While the anonymity of the page's administrators is another example of the opacity of social-media platforms, the reasons for this became clear following the supposed 'unveiling' of one of the administrators as Mandy Owens. Beginning mere hours after the release of her name, Owens received a steady stream of death threats, rape threats, and threats to her child. Owens' home address was posted online, her home vandalised, and her car tyres slashed. She lost her job. A year on, she still feared being recognised and physically harmed or verbally abused.[91] Her public naming is indicative of the extreme nature of views relating to farm murders on social media and how these can slip dangerously into real-world actions.

Why does the issue of farm murders touch such sensitive chords? Those who mobilise around the issue often present it as the latest manifestation of a history of Afrikaner victimhood. At the centre of this old narrative are the loss of the Boer Republics and the deaths of women and children in concentration camps during the South African War, as can be seen in Figures 2.4.3 and 2.4.4, which appeared on the Busting the Myth of White Genocide in SA Facebook page. Here past and present intertwine: whereas once Afrikaners were victims of the "English", now they are the victims of "black people."



Laura van der Walt
Fok off! You kill our people everyday ! What if white people killed your people everyday ????how would you feel ?
Like · Reply · 1y
13

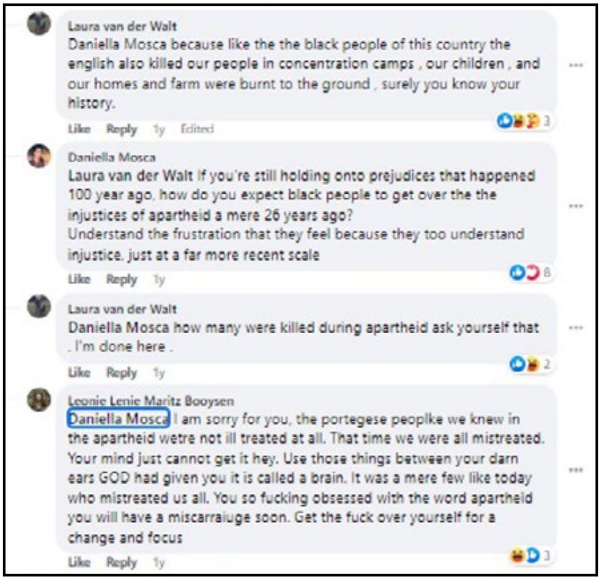
Hide 12 Replies

Ashraf Adams
Laura van der Walt I'm sorry who here killed your people?
Like · Reply · 1y
9

Laura van der Walt
Ashraf Adams black males
Like · Reply · 1y
1

Figure 2.4.3

[91] Simon Allison, 'The Facebook group taking on South Africa's white right', *Mail & Guardian*, 28 February 2020.



Laura van der Walt
Daniella Mosca because like the the black people of this country the english also killed our people in concentration camps , our children , and our homes and farm were burnt to the ground , surely you know your history.

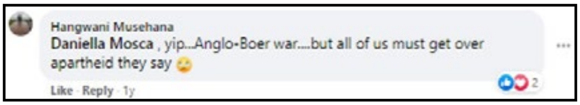
Daniella Mosca
Laura van der Walt If you're still holding onto prejudices that happened 100 year ago, how do you expect black people to get over the the injustices of apartheid a mere 26 years ago?
Understand the frustration that they feel because they too understand injustice. just at a far more recent scale

Laura van der Walt
Daniella Mosca how many were killed during apartheid ask yourself that . I'm done here .

Leonie Lenie Maritz Booysen
Daniella Mosca I am sorry for you, the portegese peoplke we knew in the apartheid wetre not ill treated at all. That time we were all mistreated. Your mind just cannot get it hey. Use those things between your darn ears GOD had given you it is called a brain. It was a mere few like today who mistreated us all. You so fucking obsessed with the word apartheid you will have a miscarraiuge soon. Get the fuck over yourself for a change and focus

Figure 2.4.4

This set of interactions not only highlights a selective mobilisation of history that was common in the content we examined, but also the constant calls to 'move on from apartheid' by users who unabashedly invoked other historical episodes. The irony did not escape a participant in this same discussion who sarcastically wrote on the thread "yip…Anglo-Boer war… but all of us must get over apartheid they say" (see Figure 2.4.5).



Hangwani Musehana
Daniella Mosca , yip...Anglo-Boer war....but all of us must get over apartheid they say 😂

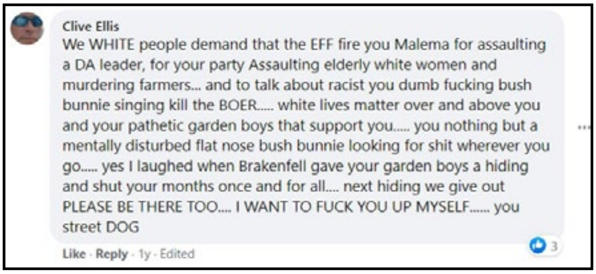
Figure 2.4.5

Sensitivity around farm murders also reflects the romanticised status of the farm in Afrikaner culture.[92] Protecting the land came to form a key feature of Afrikaner identity: protecting it from the British in the nineteenth century and protecting it from the 'rooi gevaar' and the 'swart gevaar' in the twentieth century. These fears have now coalesced around Julius Malema and the EFF. The most vocal proponent of land expropriation, Malema has called for land occupation and redistribution,

[92] Hannelie Marx Knoetze, 'Romanticising the "Boer": Narratives of White Victimhood in South African Popular Culture', in *Journal of Literary Studies,* Vol. 36, No. 4 (2020), pp. 48–69, at page 54.

since the inception of the EFF.[93] This may in part explain the level of vitriol directed at him and the EFF, as well as the ways in which the party often became the prime focus of political discussion in the content we examined.

Social media discussion of Malema and the EFF was particularly strident when it comes to the issue of 'farm murders'. Derogatory comments as well as racist tropes and stereotypes were ubiquitous (see Figure 2.4.6 for an extreme example).



Clive Ellis

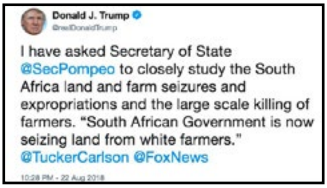
We WHITE people demand that the EFF fire you Malema for assaulting a DA leader, for your party Assaulting elderly white women and murdering farmers... and to talk about racist you dumb fucking bush bunnie singing kill the BOER..... white lives matter over and above you and your pathetic garden boys that support you..... you nothing but a mentally disturbed flat nose bush bunnie looking for shit wherever you go..... yes I laughed when Brakenfell gave your garden boys a hiding and shut your months once and for all.... next hiding we give out PLEASE BE THERE TOO.... I WANT TO FUCK YOU UP MYSELF...... you street DOG

Figure 2.4.6

Aside from racist venting of this kind, a variety of activist groups, including AfriForum, which describes itself as a civil rights organisation that mobilises to protect Afrikaner and other minority groups rights but which has often been described as a white nationalist organisation, have seen opportunity in presenting themselves as defenders of white farmers both on social media and in the real world. Retweets by and responses to Ernst Roets (the Deputy CEO of AfriForum and its public face) appeared sporadically across our dataset, suggesting the way in which such interests have sought to make hay from discussion around farm murders.[94] AfriForum has become a significant voice on the local and international stage, touring the US and UK in 2018 to call attention to 'the plight' of white farmers, which led President Trump to tweet about the issue (see Figure 2.4.7 below).

[93] In their manifesto for the 2021 local government elections, the first two points were: 1. The expropriation of South Africa's land without compensation for equal redistribution; and 2. The nationalisation of mines, banks, and other strategic sectors of the economy, without compensation. Lizeka Tandwa, 'EFF produces ambitious manifesto, promising land redistribution', *Mail & Guardian,* 26 September 2021, https://mg.co.za/politics/2021-09-26-eff-produces-ambitious-manifesto-promising-land-redistribution/

[94] This positioning of whites as a persecuted minority has been a key project of Afriforum, particularly Ernst Roets, who argues in his 2018 book *Kill the Boer: Government Complicity in South Africa's Brutal Farm Murders*, as the title suggests, that the South African government is complicit in this 'crisis'. In the book, he writes "[i]n recent years there has been a gradual increase in international reporting on farm murders, often with particular focus on the South African government's careless attitude towards the problem. Talks of a looming white genocide have also increased dramatically". Ernst Roets, *Kill the Boer: Government Complicity in South Africa's Brutal Farm Murders* (Pretoria: Kraal Uitgewers [part of the Solidarity Movement], 2018).


Donald J. Trump
@realDonaldTrump

I have asked Secretary of State @SecPompeo to closely study the South Africa land and farm seizures and expropriations and the large scale killing of farmers. "South African Government is now seizing land from white farmers." @TuckerCarlson @FoxNews

10:28 PM - 22 Aug 2018

Figure 2.4.7.[95]

Although AfriForum has stopped short of explicitly using the term "white genocide", Roets has freely associated with individuals such as Willem Petzer, who are propagators of white genocide theory and has popularised the idea that whites are a persecuted minority in South Africa.[96] Other groups have taken more extreme positions. Suidlanders – which describes itself as a civil defence organisation – have, for example, formed close links to the international far right. The group draws on the apocalyptic prophecies of Nicolaas 'Siener' van Rensburg.[97] These prophecies have been remobilised in the present and spread by social media. For example, there are numerous videos relating to the prophecies of 'Siener' on YouTube. In our own data relating to the Brackenfell High School protests, a Twitter user, for example, finds comfort in van Rensburg's prediction that the death of a black leader would be followed by violence and civil war (Figure 2.4.8, with its English translation in Figure 2.4.9). As we have seen, the anticipation of civil war was a recurrent theme in the discussions of Senekal and Brackenfell.

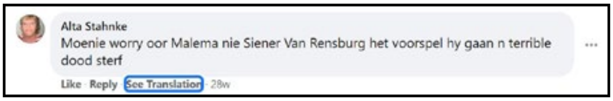

Alta Stahnke
Moenie worry oor Malema nie Siener Van Rensburg het voorspel hy gaan n terrible dood sterf

Figure 2.4.8

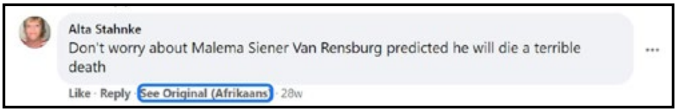

Alta Stahnke
Don't worry about Malema Siener Van Rensburg predicted he will die a terrible death

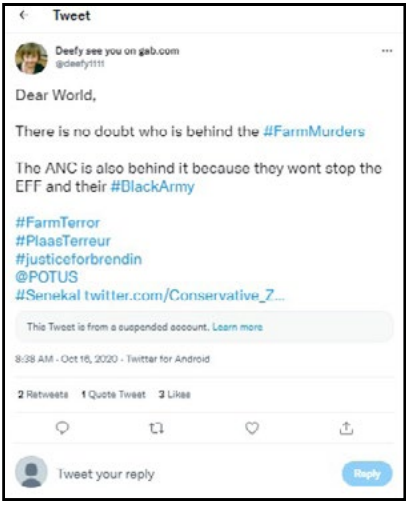
Figure 2.4.9

[95] Tweet taken from Jennifer Williams, 'Trump's tweet echoing white nationalist propaganda about South African farmers, explained', *Vox*, 23 August 2018, https://www.vox.com/policy-and-politics/2018/8/23/17772056/south-africa-trump-tweet-afriforum-white-farmers-violence

[96] Ana Deumert, 'Sensational signs, authority and the public sphere: Settler colonial rhetoric in times of change', *Journal of Sociolinguistics*, Vol. 23 (2019), pp. 467–484, at page 478.

[97] Siener van Rensburg was a Boer fighter who played a role in the South African War against the British and the failed 1914 Rebellion against the Union government after it joined World War II on the side of the Allies. His visions have been latched on to in the belief that the history of Afrikanerdom in the twentieth century was revealed to him. See, Albert Grundlingh, 'Probing the Prophet: The Psychology and Politics of the Siener van Rensburg Phenomenon' in *South African Historical Journal*, Vol. 34 (1996), pp. 225–239 and Sandra Swart, "Desperate Men': The 1914 Rebellion and the Politics of Poverty' in *South African Historical Journal*, Vol. 42, No. 1 (2000), pp. 161–175.

The links between the issue of farm killings, the EFF, and the international right can be seen in Figure 2.4.10 below, where the user in question claims that the EFF is responsible for 'farm murders' and tags a range of hashtags that peddle the narrative that farm murders are part of an orchestrated plan targeting whites in general and white Afrikaners in particular. Such claims have received a receptive audience abroad. Genocide scholar A. Dirk Moses credits this fusion of local and international interests to the capacity of the Internet to network and create communities of conspiracy theorists in a semi-clandestine way, whether within countries or across borders. Certainly, the rapid spread of imagery and information means that the case of South African farmers can quickly assume the status of fact (i.e. that they are victims of "white genocide" when the evidence suggests otherwise). Thus established in the white nationalist imagination, South Africa represents the future they fear in the former British settler colonies of Australia, Canada, and the USA, even in Europe.[98]



Deefy see you on gab.com
@deefy1111

Dear World,

There is no doubt who is behind the #FarmMurders

The ANC is also behind it because they wont stop the EFF and their #BlackArmy

#FarmTerror
#PlaasTerreur
#justiceforbrendin
@POTUS
#Senekal twitter.com/Conservative_Z...

This Tweet is from a suspended account. Learn more

8:38 AM · Oct 16, 2020 · Twitter for Android

2 Retweets 1 Quote Tweet 3 Likes

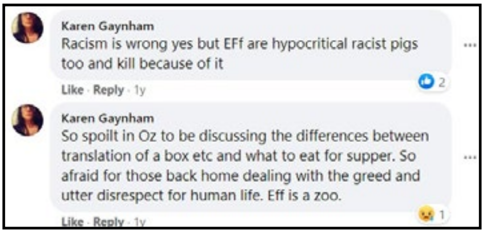
Figure 2.4.10

This process has also been aided by South African immigrants to Australia, New Zealand, and the United States who continue to post about farm murders (see, for example, Figure 2.4.11). Some have worked to popularise the notion of white genocide in South Africa. Figure 2.4.12 is particularly revealing. The tweet links to an article critiquing a protest held in New Zealand that aimed to publicise the idea that white farmers in South Africa were victims of genocidal violence due to the government's refusal to act against farm murders. The tweet was posted by the Tāmaki Anti-Fascist Action (TAFA), which describes itself as "a group mobilising against fascism and the rest of the far-right in Tāmaki Makaurau [Auckland] in

[98] A. Dirk Moses, '"White Genocide" and the Ethics of Public Analysis' in *Journal of Genocide Research*, Vol. 21, No. 2 (2019), pp. 201–213, at page 209.

Aotearoa". Here we see social media's ability to both mobilise new connections among the far right and to facilitate counter-mobilisation. We also see how the framing of Horner's death as a farm murder – rather than as one of 58 murders in South Africa on the same day – enables the issue to gain global resonance among the far right on social media.

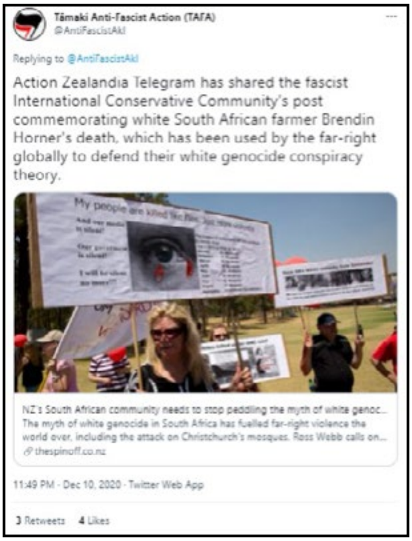
"Farm murders", in other words, mobilises the far right inside and outside South Africa.[99]



Karen Gaynham
Racism is wrong yes but EFf are hypocritical racist pigs too and kill because of it
Like · Reply · 1y 2

Karen Gaynham
So spoilt in Oz to be discussing the differences between translation of a box etc and what to eat for supper. So afraid for those back home dealing with the greed and utter disrespect for human life. Eff is a zoo.
Like · Reply · 1y 1

Figure 2.4.11



Tāmaki Anti-Fascist Action (TAFA)
@AntiFascistAkl

Replying to @AntiFascistAkl

Action Zealandia Telegram has shared the fascist International Conservative Community's post commemorating white South African farmer Brendin Horner's death, which has been used by the far-right globally to defend their white genocide conspiracy theory.

NZ's South African community needs to stop peddling the myth of white genoc...
The myth of white genocide in South Africa has fuelled far-right violence the world over, including the attack on Christchurch's mosques. Ross Webb calls on...
thespinoff.co.nz

11:49 PM · Dec 10, 2020 · Twitter Web App

3 Retweets 4 Likes

Figure 2.4.12

[99] James Pogue, 'The Myth of White Genocide: An unfinished civil war inspires a global delusion', *Harper's Magazine*, March 2019. https://harpers.org/archive/2019/03/the-myth-of-white-genocide-in-south-africa/

2.5 Presences and Absences

One of the distinguishing features of social media – discussed in detail in Section Three – is the opacity of content moderation. No case better demonstrates the hidden hand of the moderator than the relative absence of antisemitism in the content we collected relating to the 2021 Israel-Gaza conflict. Here, however, the absence of evidence is not evidence of absence. Our annotators consistently reported that important threads and conversations seemed to be 'missing'.

While we still found, during the conflict, isolated examples of antisemitism that drew on classical antisemitic tropes (see Figure 2.5.1. for the charge of deicide and Figure 2.5.2 for the charge that Jews are a cruel and heartless people), more common were debates regarding whether criticism of Zionism or Israel was antisemitic. An example of this can be seen in Figure 2.5.3. In the South African context, a comparison between Israel and the apartheid state was a recurrent feature, as will be discussed below.

Paradoxically, the relative lack of antisemitic postings relating to the 2021 Israel-Gaza conflict, as well as the presence of content pockmarked by deletions, in all likelihood did not highlight absence, but abundance. Content-moderation processes had identified and eliminated antisemitic content before we assembled our dataset. We were thus left with a breadcrumb trail of vestigial clues. This does raise the question of why content moderation was so much more successful when it came to the 2021 Israel-Gaza conflict than to the flashpoints we have described so far?

The global nature of antisemitism (and the global nature of social media responses to the 2021 Israel-Gaza conflict) may have meant that antisemitic tropes and stereotypes were more easily recognised by users and commercial content moderators. This in turn likely meant that algorithmic processes (while not infallible) had larger datasets to be trained with and in turn were better armed to remove large amounts of problematic material.

In addition, social-media platforms focused energy and attention on content moderation during the conflict. Facebook, for example, rapidly added new terms flagged for algorithmic content moderation and set up a 24/7 "special operations centre" to deal with content relating to the conflict. Other moderation efforts included increased use of geoblocking, whereby social-media companies targeted the geographical location of content to help their moderation efforts. There were also reports of the blocking of hashtags, most controversially Instagram's removal and blocking of posts with the hashtag for the Al-Aqsa Mosque after its moderation system mistakenly deemed the building a terrorist organisation.[100] Here we see the opacity of the content-moderation process, as well as the tensions between

[100] Adam Smith, 'Palestinians' Digital Rights 'Violated' by Censorship on Facebook, Twitter, and Instagram, New Report Claims', *The Independent*, 21 May 2021. https://www.independent.

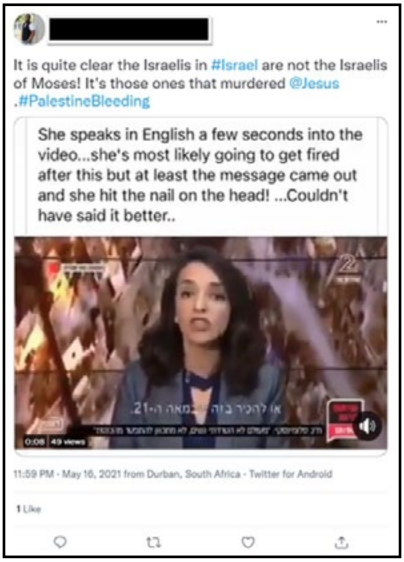
protecting free speech and removing hate speech. During the conflict, critics claimed that these processes were censoring pro-Palestinian content.[101]

As we will see, Twitter and Facebook are much less successful in moderating content that is particular to the South African context. Though we found little content that was overtly antisemitic connected with the other flashpoints, these episodes rarely lent themselves to the mobilisation of antisemitic discourse. This, however, does not indicate that antisemitic content is rare on South African social media. To give just one extreme example, in May 2020 the Pretoria 'fitness queen' and influencer Simone Kriel posted an antisemitic rant on Instagram claiming that "The f***n Jews are greedy as f**k and they will wage war against countries and races, based on lies and deception to get what they want". She continued by claiming that "It was the Jews that bombed, raped, sodomised and burned all people in Germany alive. Hitler innocent. Our history has been twisted to favour the Jews without question".

A clear example of the derogatory use of the term 'Zionist' came in a thread relating to the removal of Clover products from a Spar store in District Six, Cape Town following the purchase of Clover by an Israeli company. In response to criticism of this boycott on 29 May 2021 on Twitter, a poster wrote that those against the protest were apartheid supporters and that the user in question was a "Zionest shit stain". While this account has since been suspended, the reason for the suspension is unclear. Did the post remain on Twitter long enough for our extraction to take place because they had misspelt 'Zionist' (perhaps purposefully)? Was the post not flagged by others? Or was the post indeed flagged but Twitter decided it did not warrant removal? This example also signals the broader opacity regarding the content-moderation policies across platforms. Twitter, Facebook and TikTok pursue very different content-moderation policies.

co.uk/life-style/gadgets-and-tech/palestine-israel-censorship-facebook-twitter-instagram-7amleh-b1851328.html?amp

[101] See Ibid.; Kelly Lewis, 'Social media platforms are complicit in censoring Palestinian voices', *The Conversation*, 24 May 2021. https://theconversation.com/social-media-platforms-are-complicit-in-censoring-palestinian-voices-161094



[redacted]

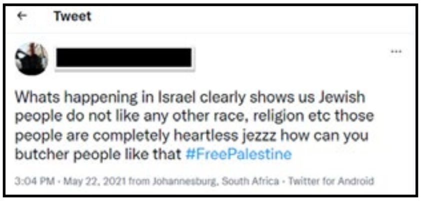
It is quite clear the Israelis in #Israel are not the Israelis of Moses! It's those ones that murdered @Jesus .#PalestineBleeding

She speaks in English a few seconds into the video...she's most likely going to get fired after this but at least the message came out and she hit the nail on the head! ...Couldn't have said it better..

11:59 PM · May 16, 2021 from Durban, South Africa · Twitter for Android

1 Like

Figure 2.5.1



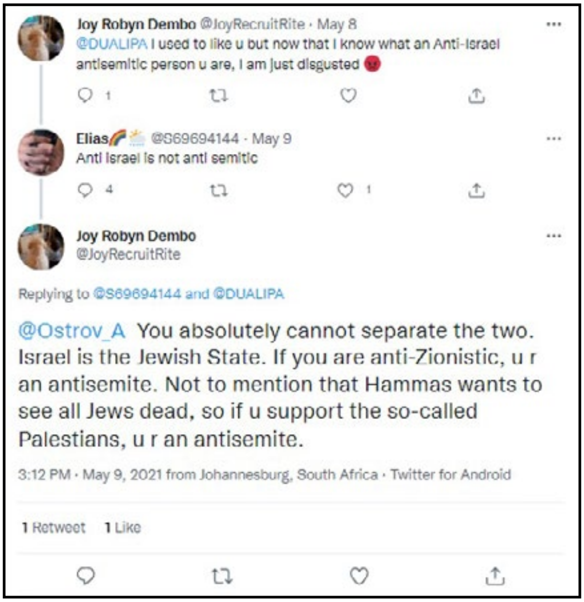
Tweet

[redacted]

Whats happening in Israel clearly shows us Jewish people do not like any other race, religion etc those people are completely heartless jezzz how can you butcher people like that #FreePalestine

3:04 PM · May 22, 2021 from Johannesburg, South Africa · Twitter for Android

Figure 2.5.2



Joy Robyn Dembo @JoyRecruitRite · May 8

@DUALIPA I used to like u but now that I know what an Anti-Israel antisemitic person u are, I am just disgusted 😡

1

Elias🌈🌦️ @S69694144 · May 9

Anti Israel is not anti semitic

4 1

Joy Robyn Dembo
@JoyRecruitRite

Replying to @S69694144 and @DUALIPA

@Ostrov_A You absolutely cannot separate the two. Israel is the Jewish State. If you are anti-Zionistic, u r an antisemite. Not to mention that Hammas wants to see all Jews dead, so if u support the so-called Palestians, u r an antisemite.

3:12 PM · May 9, 2021 from Johannesburg, South Africa · Twitter for Android

1 Retweet 1 Like

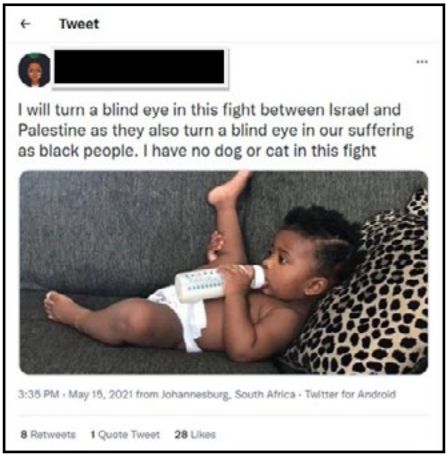
Figure 2.5.3

What direction did discussion of the 2021 Israel-Gaza conflict take on social media in South Africa?

Soon after the outbreak of violence in Gaza in May 2021, various high-level members of the ANC and South African government criticised the actions of Israel. For example, Jessie Duarte, the ANC Deputy Secretary-General led a picket of the Israeli embassy in Pretoria on 25 May 2021 during which she stated that genocide was occurring in Palestine and that Israel was becoming a global imperialist that would soon be a threat to African land. In the days following, Naledi Pandor compared the experiences of Palestinians to being “trapped in an apartheid manner“ and President Cyril Ramaphosa highlighted the Palestinian people’s right to self-determination.

Online discussion about the conflict soon veered towards criticism of the ANC and the government. This reflected a broader trend: the refraction of discussion of international events through the lens of national politics. Responses to the 2021 Israel-Gaza conflict coalesced around local issues. Three themes appeared consistently.

The first theme was a refusal by users to engage with the event (see Figure 2.5.4, for example). The second promoted the view that the South African government should not get involved in the conflict as South Africa has problems enough of its own (see, for example, Figures 2.5.5 and 2.5.6). These views sometimes spilt into crude racial stereotypes (see Figure 2.5.7).



Tweet

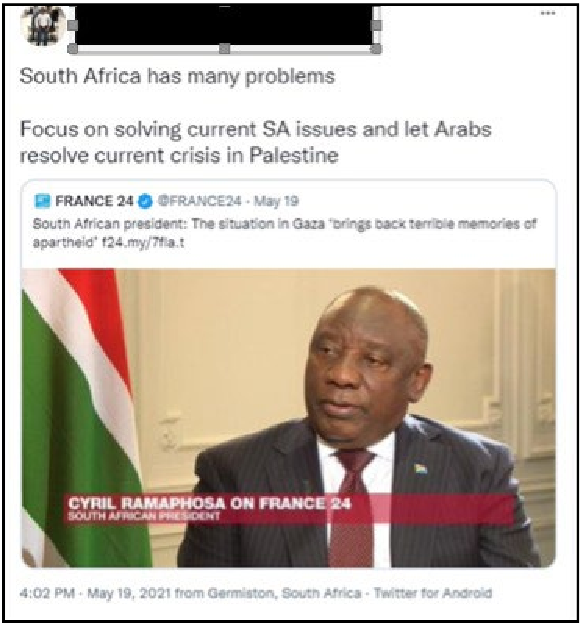
[redacted]

I will turn a blind eye in this fight between Israel and Palestine as they also turn a blind eye in our suffering as black people. I have no dog or cat in this fight

3:35 PM · May 15, 2021 from Johannesburg, South Africa · Twitter for Android

8 Retweets 1 Quote Tweet 28 Likes

Figure 2.5.4



[redacted]

South Africa has many problems

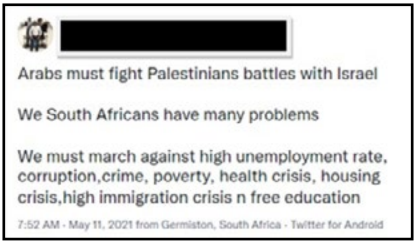
Focus on solving current SA issues and let Arabs resolve current crisis in Palestine

FRANCE 24 @FRANCE24 · May 19

South African president: The situation in Gaza 'brings back terrible memories of apartheid' f24.my/7fIa.t

4:02 PM · May 19, 2021 from Germiston, South Africa · Twitter for Android

Figure 2.5.5



[redacted]

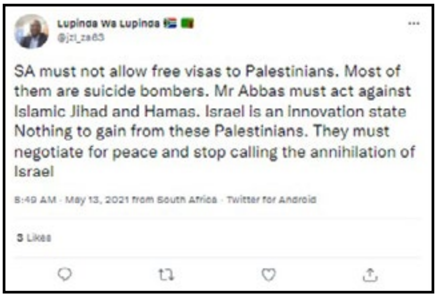
Arabs must fight Palestinians battles with Israel

We South Africans have many problems

We must march against high unemployment rate, corruption,crime, poverty, health crisis, housing crisis,high immigration crisis n free education

7:52 AM · May 11, 2021 from Germiston, South Africa · Twitter for Android

Figure 2.5.6



Lupinda Wa Lupinda 🇿🇦 🇿🇲
@jzi_za83

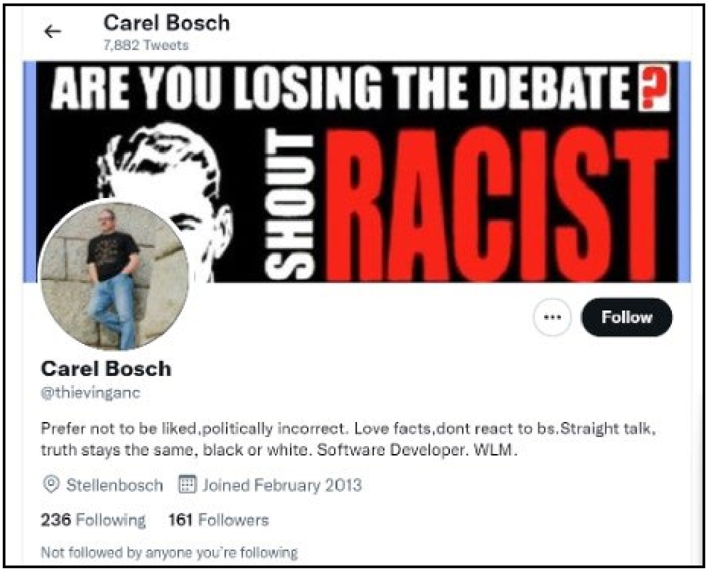
SA must not allow free visas to Palestinians. Most of them are suicide bombers. Mr Abbas must act against Islamic Jihad and Hamas. Israel is an innovation state Nothing to gain from these Palestinians. They must negotiate for peace and stop calling the annihilation of Israel

8:49 AM · May 13, 2021 from South Africa · Twitter for Android

3 Likes

Figure 2.5.7

A more vitriolic version of this view involved the claim that the government itself was incompetent and thus had no right to criticise others. An extreme example of this was seen in a response to a tweet by Carl Niehaus dated 11 May 2021 that contained a video of the storming of the Al Aqsa Mosque by Israeli soldiers, which Niehaus referred to as a "Scandalous act of repression by the apartheid racist State of Israel." A user with the handle @thievinganc responded with, "Go Israel, flatten the Palestine hell hole of terrorists. Can we send you some political opportunists, observers, before you do?". Racial slurs were repeated throughout this user's comments regarding Palestinians and black Africans. The user's profile (Figure 2.5.8) refers to WLM (White Lives Matter) and suggests that his views on Israel reflect his broader white supremacist worldview. Others engaged in similar rhetoric that also turned an international event into an opportunity to rant about local grievances (see, for example, Figure 2.5.9).



Carel Bosch
7,882 Tweets

Carel Bosch
@thievinganc

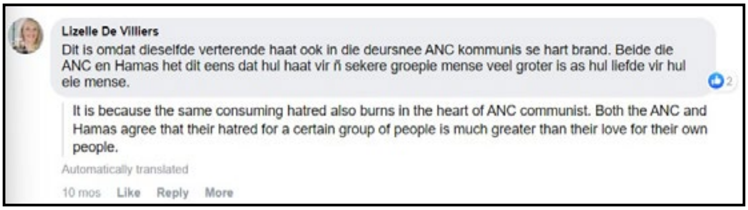
Prefer not to be liked,politically incorrect. Love facts,dont react to bs.Straight talk, truth stays the same, black or white. Software Developer. WLM.

Stellenbosch Joined February 2013

236 Following 161 Followers

Not followed by anyone you're following

Figure 2.5.8



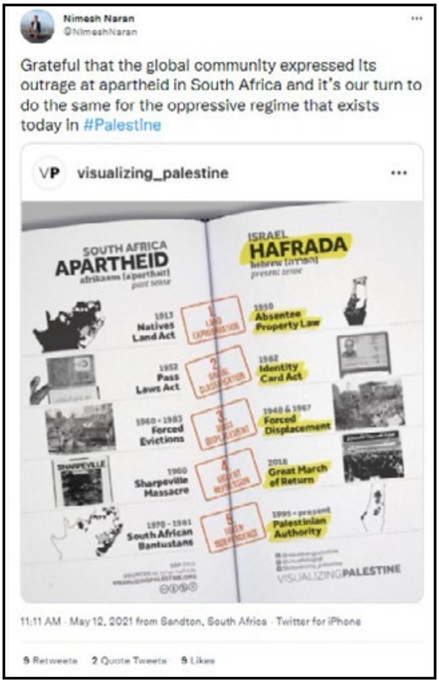
Lizelle De Villiers
Dit is omdat dieselfde verterende haat ook in die deursnee ANC kommunis se hart brand. Beide die ANC en Hamas het dit eens dat hul haat vir ñ sekere groepie mense veel groter is as hul liefde vir hul eie mense.

It is because the same consuming hatred also burns in the heart of ANC communist. Both the ANC and Hamas agree that their hatred for a certain group of people is much greater than their love for their own people.

Automatically translated

Figure 2.5.9

This merging of local concerns with international events was particularly pronounced on Facebook, where the charge that Israel is akin to apartheid South Africa was ubiquitous (see, for example, Figures 2.5.10 and 2.5.11). This was the third theme that emerged in the content we analysed. Those who invoked this comparison sought to draw on the past as a form of validation – if the apartheid state was unjust, immoral, and deserving of dismantling, then so is Israel – and to make the events in Palestine relatable to a broader South African audience. In line with the tendency of users to localise the global, the comparison often evoked responses that focused on South Africa, not Israel. For example, in response to Naledi Pandor's statement that Palestinians are "trapped in an apartheid manner", some ostensibly white and black users carped that Pandor should instead focus on issues at home (Figures 2.5.12 and 2.5.13).



Nimesh Naran
@NimeshNaran

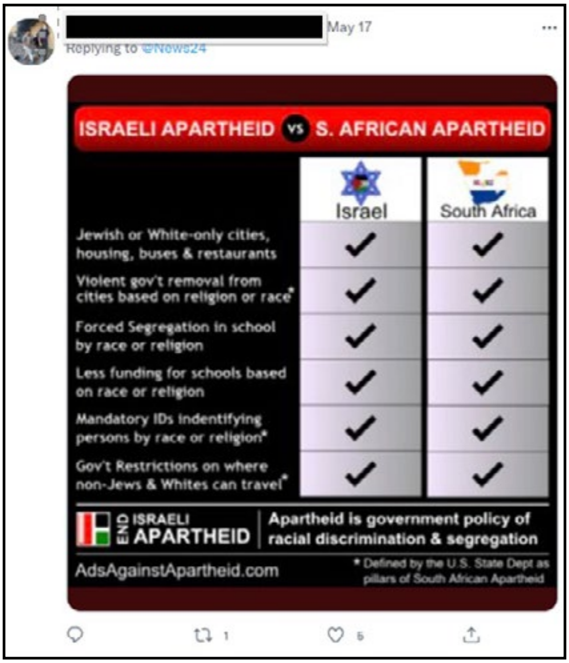
Grateful that the global community expressed its outrage at apartheid in South Africa and it's our turn to do the same for the oppressive regime that exists today in #Palestine

visualizing_palestine

11:11 AM · May 12, 2021 from Sandton, South Africa · Twitter for iPhone

5 Retweets 2 Quote Tweets 5 Likes

Figure 2.5.10



[redacted] May 17

Replying to @News24

ISRAELI APARTHEID VS S. AFRICAN APARTHEID

	Israel	South Africa
Jewish or White-only cities, housing, buses & restaurants	✓	✓
Violent gov't removal from cities based on religion or race*	✓	✓
Forced Segregation in school by race or religion	✓	✓
Less funding for schools based on race or religion	✓	✓
Mandatory IDs indentifying persons by race or religion*	✓	✓
Gov't Restrictions on where non-Jews & Whites can travel*	✓	✓

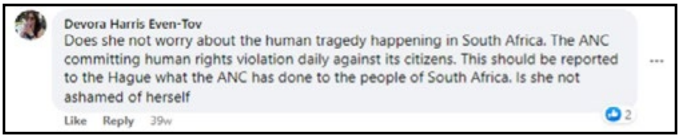
END ISRAELI APARTHEID

Apartheid is government policy of racial discrimination & segregation

AdsAgainstApartheid.com

* Defined by the U.S. State Dept as pillars of South African Apartheid

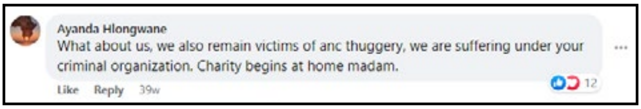
Figure 2.5.11



Devora Harris Even-Tov

Does she not worry about the human tragedy happening in South Africa. The ANC committing human rights violation daily against its citizens. This should be reported to the Hague what the ANC has done to the people of South Africa. Is she not ashamed of herself

Figure 2.5.12



Ayanda Hlongwane

What about us, we also remain victims of anc thuggery, we are suffering under your criminal organization. Charity begins at home madam.

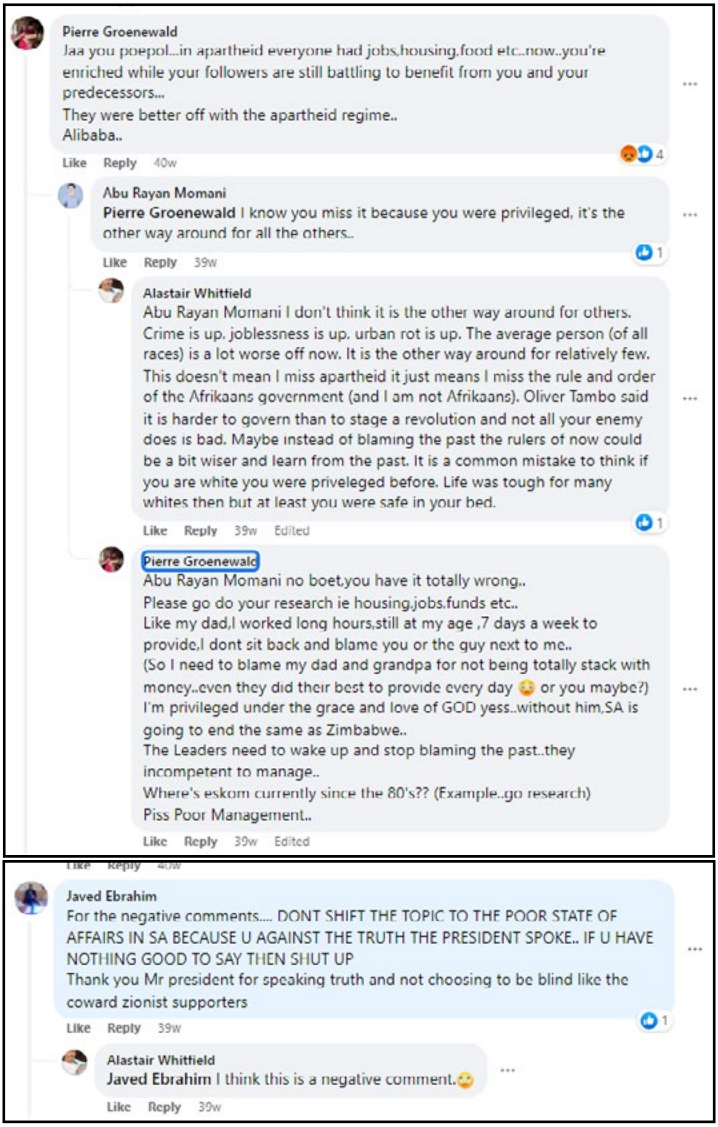
Figure 2.5.13

Figure 2.5.14 reflects what was purportedly a discussion of Pandor's words. Here we see the apotheosis of this tendency to find the local in the global, as well as the appearance of stock tropes that recurred across all of the flashpoints we examined: nostalgia for the past where things were supposedly "better for everyone"; the notion that South African leaders (and by implication all non-whites) are incompetent and incapable of rule; and the idea that "South Africa will end up like Zimbabwe“. These comments appear on a thread engaged with a completely unrelated topic: the Israel-Gaza conflict.

Attempts to shift the terms of the debate are a constant (and often successful) feature of discussion on social media. In the above case, a user steps in to make a vain attempt to bring the conversation back to the matter at hand by stating: “DON'T SHIFT THE TOPIC TO THE POOR STATE OF AFFAIRS IN SA BECAUSE U AGAINST THE TRUTH THE PRESIDENT SPOKE..IF YOU HAVE NOTHING GOOD TO SAY THEN SHUT UP“. The user ends their comment by referring to those not speaking out against Israel as “coward Zionist supporters“.

The use of capitalisation intended to shout at other users to ‘keep quiet’, ‘shut up’ or ‘go away’ was a common feature across all the flashpoints we examined, as was the alternate ‘mxm’ (the sound made by kissing your teeth, which in South Africa signifies dismissing what someone else has said). In this instance the nature of the discussion, and its conclusion, suggests the ways in which these platforms act more as a site for dismissing the comments of others than as forums for engagement with them. In this case, the shouting at another user to shut up was met by a rather forlorn “I think this is a negative comment“, which brought the ‘debate’ to an end.

As this and other cases demonstrate, social media discussions of the Israel-Gaza conflict were liable to quickly degenerate into crudity. In some measure this may have reflected a dynamic embedded within the Israel-apartheid analogy. When this analogy was used, it repeatedly generated defensiveness among some ostensibly white social media users who took this and other mentions of apartheid as challenges to their status and belonging as South Africans. The analogy, in other words, exacerbated an existing tendency that these users shared with many others on social media in South Africa: a desire, first and foremost, to fight about local political issues.



Pierre Groenewald
Jaa you poepol...in apartheid everyone had jobs,housing,food etc..now..you're enriched while your followers are still battling to benefit from you and your predecessors...
They were better off with the apartheid regime..
Alibaba..

Like Reply 40w 4

Abu Rayan Momani
Pierre Groenewald I know you miss it because you were privileged, it's the other way around for all the others..

Like Reply 39w 1

Alastair Whitfield
Abu Rayan Momani I don't think it is the other way around for others. Crime is up. joblessness is up. urban rot is up. The average person (of all races) is a lot worse off now. It is the other way around for relatively few. This doesn't mean I miss apartheid it just means I miss the rule and order of the Afrikaans government (and I am not Afrikaans). Oliver Tambo said it is harder to govern than to stage a revolution and not all your enemy does is bad. Maybe instead of blaming the past the rulers of now could be a bit wiser and learn from the past. It is a common mistake to think if you are white you were priveleged before. Life was tough for many whites then but at least you were safe in your bed.

Like Reply 39w Edited 1

Pierre Groenewald
Abu Rayan Momani no boet,you have it totally wrong..
Please go do your research ie housing,jobs,funds etc..
Like my dad,I worked long hours,still at my age ,7 days a week to provide,I dont sit back and blame you or the guy next to me..
(So I need to blame my dad and grandpa for not being totally stack with money..even they did their best to provide every day 😓 or you maybe?)
I'm privileged under the grace and love of GOD yess..without him,SA is going to end the same as Zimbabwe..
The Leaders need to wake up and stop blaming the past..they incompetent to manage..
Where's eskom currently since the 80's?? (Example..go research)
Piss Poor Management..

Like Reply 39w Edited

Like Reply 40w

Javed Ebrahim
For the negative comments.... DONT SHIFT THE TOPIC TO THE POOR STATE OF AFFAIRS IN SA BECAUSE U AGAINST THE TRUTH THE PRESIDENT SPOKE.. IF U HAVE NOTHING GOOD TO SAY THEN SHUT UP
Thank you Mr president for speaking truth and not choosing to be blind like the coward zionist supporters

Like Reply 39w 1

Alastair Whitfield
Javed Ebrahim I think this is a negative comment.🙄

Like Reply 39w

Figure 2.5.14

2.6 The Power of Analogy: Hitler's Hold on the South African Imagination

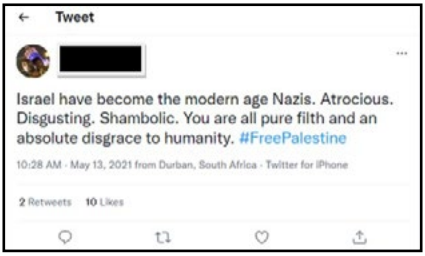
One of the most striking features of our analysis of social media in South Africa was the ubiquity of invocations of the Holocaust, Hitler, and Nazism. The ready resort to these analogies provides some indication of the tone and tenor of online discussion – often vitriolic, intemperate, and drawn to extremes – and reveals the symbolic place of the Holocaust, Hitler, and Nazism within the South African imagination as the ultimate embodiment of evil.[102] Though invocations of this triumvirate are typically intended by social media users to act as an unimpeachable moral shorthand – to make definitive statements about what is right and wrong – evidence from social-media content produced during the Senekal and Brackenfell protests instead demonstrates how little agreement there is on social media about what exactly is right and wrong in South Africa. This in part reflects the polarised and fractured political landscape in South Africa, but also the hyper-stimulated, funhouse-mirror character of social media discussion.

Unsurprisingly, the Hitler/Nazi/Holocaust analogy featured in the discussion of the 2021 Israel-Gaza conflict. Here it followed an international playbook; there was little that was distinct or particular in the ways that it was used in the South African social media environment. As elsewhere in the world, the analogy was typically used to accuse Israel of genocidal violence in Palestine. One of the more extreme examples of this can be seen in Figure 2.6.1. An interesting semantic slip can be seen here in the shift from critiquing Israel to 'you all'. Responsibility and blame have expanded from the actions of Israel to an undefined 'you all', which may refer to Jews or perhaps to all supporters of Israel. More common was the use of images that were reproduced on various occasions across our datasets (see, for example, Figure 2.6.2 and Figure 2.6.3). As we will see below, these were often accompanied by text to form 'image macros' (images with superimposed text). Image macros are designed to be shared across social-media platforms and are more difficult to filter for hate speech. Because of the format of such images, it is almost always impossible to tell the origins or source of the image. They act as a form of repetitive messaging, making them particularly effective pieces of propaganda.[103] These

[102] Jonathan Jansen has also identified this fixation. See Jonathan Jansen, 'This is why Hitler-hailing attention seekers are so dangerous in SA', TimesLIVE, 14 April 2021.

[103] See Jessica Ann Barraclough, Facebook's 'White Genocide' Problem: A Sociotechnical Exploration of Problematic Information, Shareability, and Social Correction in a South African Context (Masters Dissertation, University of Cape Town, 2020), at page 26 and Eline Zenner and Dirk Geeraerts, 'One does not simply process memes: Image macros as multimodal constructions' in Esme Winter-Froemel and Verena Thaler (eds.), *Cultures and Traditions of Wordplay and Wordplay Research* (Berlin/Boston: Walter de Gruyter GmbH, 2018), at page 167.

images also often elicit an affective response which allows for greater spread and algorithmic promotion.[104]



Tweet

[redacted]

Israel have become the modern age Nazis. Atrocious. Disgusting. Shambolic. You are all pure filth and an absolute disgrace to humanity. #FreePalestine

10:28 AM · May 13, 2021 from Durban, South Africa · Twitter for iPhone

2 Retweets 10 Likes

Figure 2.6.1

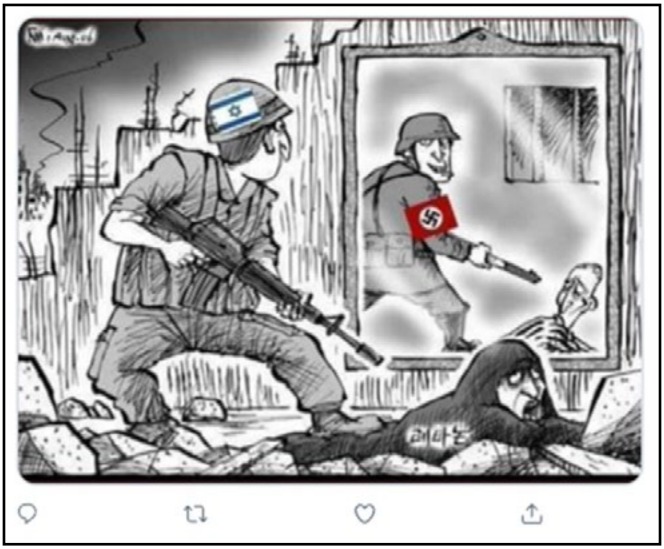
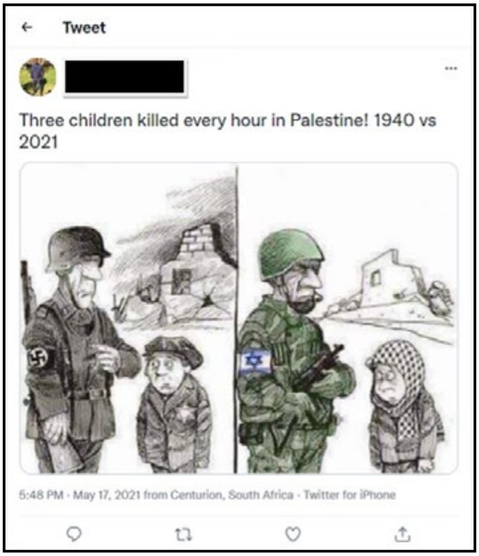


Figure 2.6.2

[104] Jessica Ann Barraclough, Facebook's 'White Genocide' Problem: A Sociotechnical Exploration of Problematic Information, Shareability, and Social Correction in a South African Context (Masters Dissertation, University of Cape Town, 2020), at page 90.



Tweet

[redacted]

Three children killed every hour in Palestine! 1940 vs 2021

5:48 PM · May 17, 2021 from Centurion, South Africa · Twitter for iPhone

Figure 2.6.3

By contrast, the invocation of the Holocaust, Hitler, and Nazism took on a more distinctive form when applied in discussions relating to South African political life. Analogies were a prominent feature of Senekal and Brackenfell content, particularly image macros that linked Julius Malema to Hitler. This was done by superimposing an image of Malema's face onto a stock image of an SS uniform. To heighten the link, Hitler's distinctive 'toothbrush' moustache has also been superimposed onto the image (see Figure 2.6.4). The stock image is readily available in many meme generators – free online image makers that let users add custom resizable text, images, and more to various templates (see, for example, Figures 2.6.5 and 2.6.6).

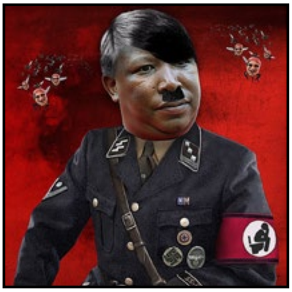


Figure 2.6.4

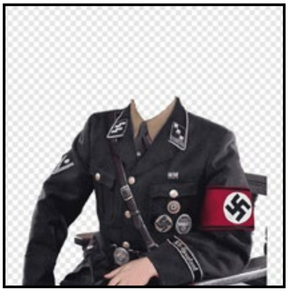


Figure 2.6.5

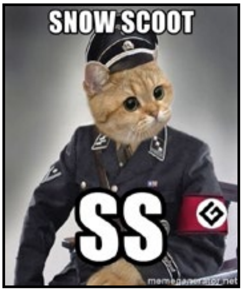
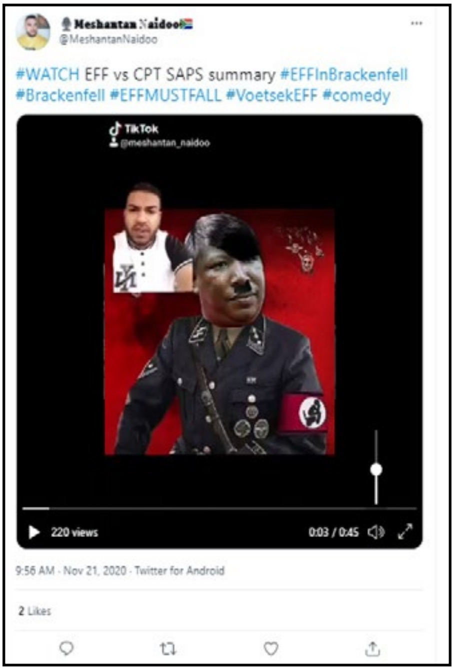


Figure 2.6.6

The image in Figure 2.6.4 serves the double purpose of portraying Malema as a genocidal tyrant in waiting and as a comedic caricature to be ridiculed. The image thus embodied fear that the EFF's policies and actions would destroy South Africa and a belief that Malema and his lieutenants could be laughed off as buffoonish populists with little real substance.

This repurposing of the Hitler analogy for the South African context via a stock meme image highlights the easy reproducibility of such images: they appear across Facebook and Twitter. In Figure 2.6.7, we see how such tropes travel across platforms; here Meshantan Naidoo (a South African comedian) used the image as the backdrop to a video clip containing a satirical summary of the protests in Brackenfell, which was initially uploaded on TikTok. He then shared the video on his Twitter account and on his YouTube account. This once again highlights that social-media platforms (as argued in Section One) need to be seen as part of a social-media ecosystem where information easily transitions across formats and platforms. This means that material removed on one platform can still exist unscathed on other platforms. On YouTube, the option exists for users to share material on 13 other online platforms.



Meshantan Naidoo
@MeshantanNaidoo

#WATCH EFF vs CPT SAPS summary #EFFInBrackenfell #Brackenfell #EFFMUSTFALL #VoetsekEFF #comedy

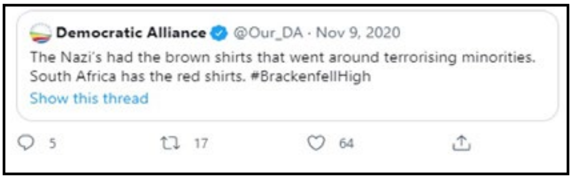
9:56 AM · Nov 21, 2020 · Twitter for Android

2 Likes

Figure 2.6.7

The comparison of Malema to Hitler is hardly new. As early as 2014, the ANC secretary-general Gwede Mantashe argued that “This movement [the EFF] uses uniform to mobilise in the same way Hitler used brown shirts in the 1930s“.[105] This was parroted by numerous other members of the ANC, including Buti Manamela who, during a parliamentary debate on the presidency budget vote in 2014, compared Malema and Hitler.[106] South African Communist Party deputy general secretary and Deputy Public Works Minister Jeremy Cronin similarly wrote in an SACP newsletter how a small number of elected Nazi Party members were able to disrupt the work of the Reichstag, implying that Malema and the EFF were following in their footsteps.[107]

The Democratic Alliance took up this theme when EFF protesters marched in Brackenfell (see Figure 2.6.8). Unlike the rhetoric of the ANC and SACP, which framed the EFF as a threat to parliamentary democracy, the DA here attempted to frame the EFF and Malema as thugs. Figure 2.6.9 frames the EFF as fascists; this tweet generated a range of responses, some very critical (see Figures 2.6.10 and 2.6.11).


Democratic Alliance @Our_DA · Nov 9, 2020

The Nazi's had the brown shirts that went around terrorising minorities. South Africa has the red shirts. #BrackenfellHigh

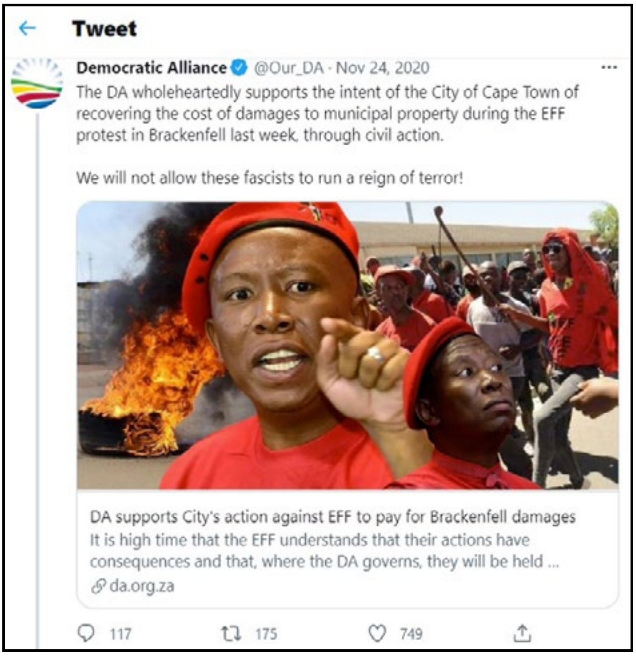
Show this thread

Figure 2.6.8

[105] ‘Alliance sees the ghost of Hitler in Malema’, News24, 16 March 2015. https://www.news24.com/news24/alliance-sees-the-ghost-of-hitler-in-malema-20150429

[106] Manamela stated: “In the 1920s, there was an incident in Germany. A young man, who was supposed to be at the helm of the country’s political elite, was sidelined and, ultimately, elbowed out of the limelight of German political elitism, dismissed as unstable and politically immature. Not to be deterred from what he believed was a calling - a sad date with political destiny - he started mobilising others who were moved by his rhetoric and stood to benefit from his ascension to power [...] Hitler, for instance, declared his ideology as national socialism. It sounded nice...” The full transcript of this Parliamentary debate can be found at https://www.politicsweb.co.za/documents/malema-adolf-hitler-has-come-back-from-the-dead--b

[107] Cronin concluded: “There are troubling additional historical echoes in our present – the cult of a megalomaniac personality with an oratorical gift; militaristic pretensions; a demagogic populism that mobilises on the basis of grievance and victimhood”. Jeremy Cronin, ‘Legislative disruptions: From the Nazis to the EFF’, 20 February 2015. https://www.politicsweb.co.za/opinion/legislative-disruptions-from-the-nazis-to-the-eff-



Tweet

Democratic Alliance @Our_DA · Nov 24, 2020

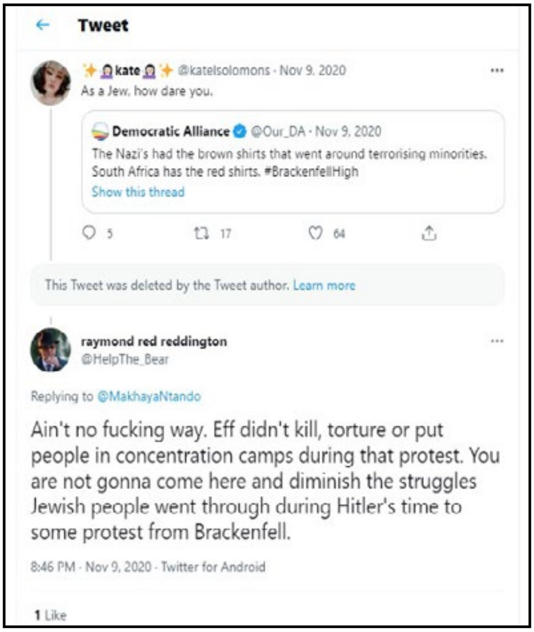
The DA wholeheartedly supports the intent of the City of Cape Town of recovering the cost of damages to municipal property during the EFF protest in Brackenfell last week, through civil action.

We will not allow these fascists to run a reign of terror!

DA supports City's action against EFF to pay for Brackenfell damages
It is high time that the EFF understands that their actions have consequences and that, where the DA governs, they will be held ...
da.org.za

117 · 175 · 749

Figure 2.6.9



Tweet

✨👩🏻kate👩🏻✨ @katelsolomons · Nov 9, 2020

As a Jew, how dare you.

> **Democratic Alliance** @Our_DA · Nov 9, 2020
> The Nazi's had the brown shirts that went around terrorising minorities. South Africa has the red shirts. #BrackenfellHigh
> Show this thread

5 · 17 · 64

This Tweet was deleted by the Tweet author. Learn more

raymond red reddington
@HelpThe_Bear

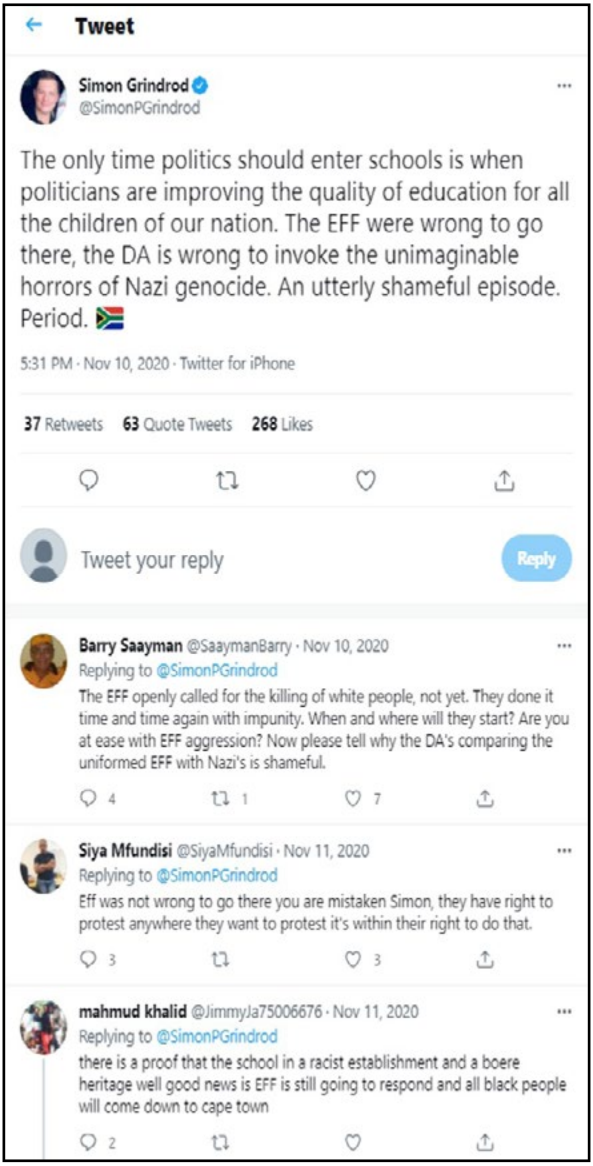
Replying to @MakhayaNtando

Ain't no fucking way. Eff didn't kill, torture or put people in concentration camps during that protest. You are not gonna come here and diminish the struggles Jewish people went through during Hitler's time to some protest from Brackenfell.

8:46 PM · Nov 9, 2020 · Twitter for Android

1 Like

Figure 2.6.10



Simon Grindrod @SimonPGrindrod

The only time politics should enter schools is when politicians are improving the quality of education for all the children of our nation. The EFF were wrong to go there, the DA is wrong to invoke the unimaginable horrors of Nazi genocide. An utterly shameful episode. Period. 🇿🇦

5:31 PM · Nov 10, 2020 · Twitter for iPhone

37 Retweets 63 Quote Tweets 268 Likes

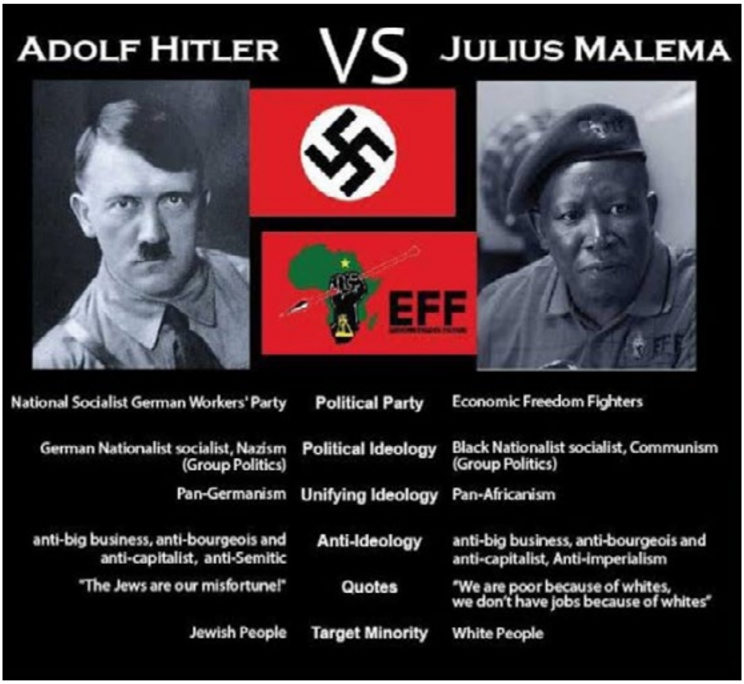
Barry Saayman @SaaymanBarry · Nov 10, 2020
Replying to @SimonPGrindrod
The EFF openly called for the killing of white people, not yet. They done it time and time again with impunity. When and where will they start? Are you at ease with EFF aggression? Now please tell why the DA's comparing the uniformed EFF with Nazi's is shameful.
4 1 7

Siya Mfundisi @SiyaMfundisi · Nov 11, 2020
Replying to @SimonPGrindrod
Eff was not wrong to go there you are mistaken Simon, they have right to protest anywhere they want to protest it's within their right to do that.
3 3

mahmud khalid @JimmyJa75006676 · Nov 11, 2020
Replying to @SimonPGrindrod
there is a proof that the school in a racist establishment and a boere heritage well good news is EFF is still going to respond and all black people will come down to cape town
2

Figure 2.6.11

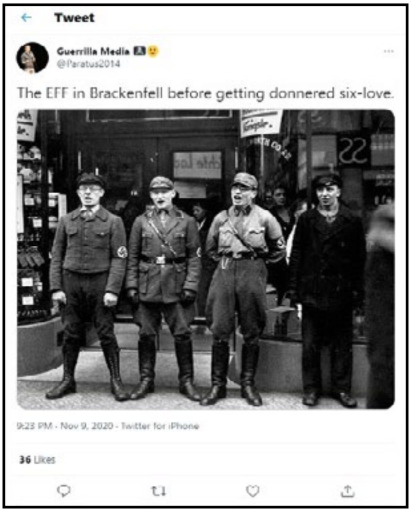
Although not as prevalent on Facebook as on Twitter, the Malema/Hitler and EFF/Brownshirt analogy still made regular appearances. When this comparison was originally invoked by the ANC, the threat was framed as that of a fringe party with a charismatic leader using the language of populism to undermine democracy. Since then, the EFF has become firmly entrenched in the South African political landscape and has made significant inroads into the ANC's share of the vote. As the EFF has grown, the Nazi analogies directed at the party and its leader have taken on new forms. Typical is Figure 2.6.12, which appeared on multiple occasions.



ADOLF HITLER	VS	JULIUS MALEMA
National Socialist German Workers' Party	Political Party	Economic Freedom Fighters
German Nationalist socialist, Nazism (Group Politics)	Political Ideology	Black Nationalist socialist, Communism (Group Politics)
Pan-Germanism	Unifying Ideology	Pan-Africanism
anti-big business, anti-bourgeois and anti-capitalist, anti-Semitic	Anti-Ideology	anti-big business, anti-bourgeois and anti-capitalist, Anti-imperialism
"The Jews are our misfortune!"	Quotes	"We are poor because of whites, we don't have jobs because of whites"
Jewish People	Target Minority	White People

Figure 2.6.12

This post also reflects the racialisation of politics on social media in South Africa. Here the post proposes that the key aim of the EFF is the destruction of white people and in doing so victimises whites in much the same manner as the Nazi party targeted Jews. This analogy between the EFF and Nazism has proven popular with those on the far right. An example of this can be seen in Figure 2.6.13 below. The user in question, whose profile page can be seen in Figure 2.6.14, makes the link here between the Nazis and the EFF, but also uses the opportunity to post material regarding gun ownership along with claims that the government does not protect its white citizens (see Figure 2.6.15).



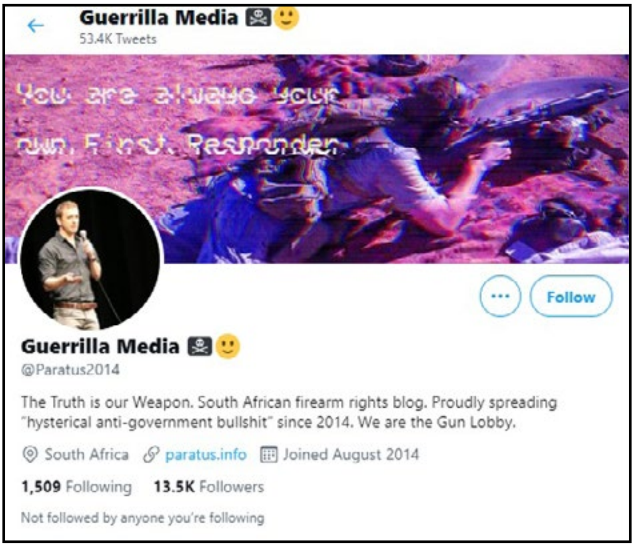
Guerrilla Media 🏴‍☠️🙂
@Paratus2014

The EFF in Brackenfell before getting donnered six-love.

9:23 PM · Nov 9, 2020 · Twitter for iPhone

36 Likes

Figure 2.6.13



Guerrilla Media 🏴‍☠️🙂
@Paratus2014

The Truth is our Weapon. South African firearm rights blog. Proudly spreading "hysterical anti-government bullshit" since 2014. We are the Gun Lobby.

South Africa paratus.info Joined August 2014

1,509 Following **13.5K** Followers

Not followed by anyone you're following

Figure 2.6.14

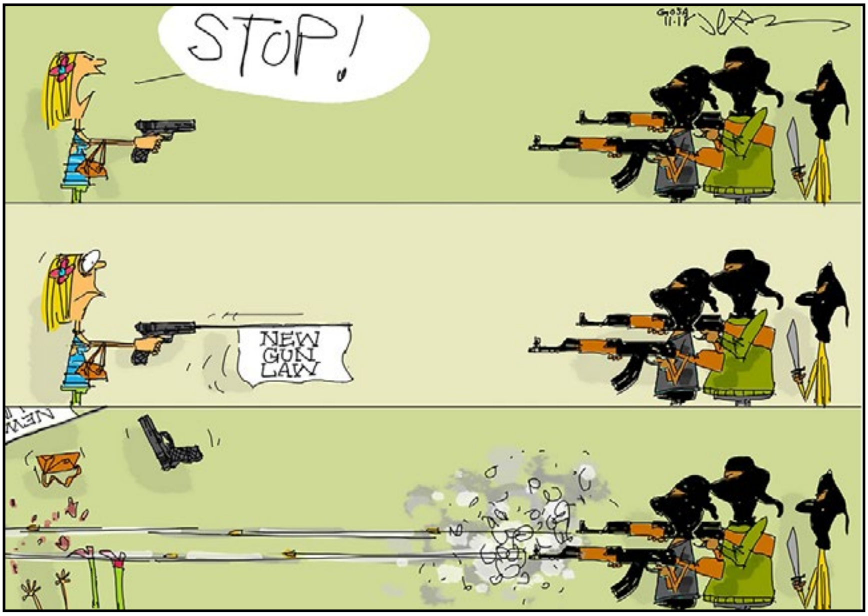
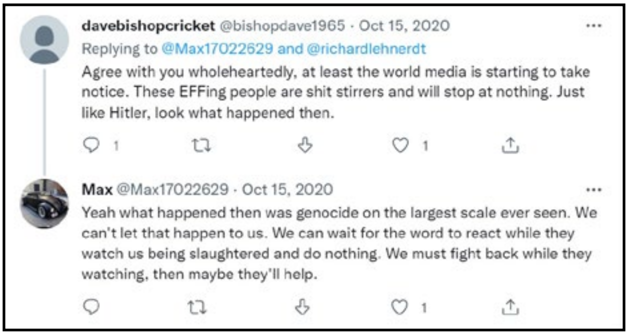


Figure 2.6.15

Fears of black majority rule, farm murders, and white genocide run together in the far-right discourse on social media. Again and again, the Holocaust is invoked as a warning against neglecting to “deal“ with “shit stirrers“ (see Figure 2.6.16). Other users posted regularly to warn of imminent genocide. The implication is clear: how can you sit idle when Malema may be “already planning where to put his concentration camps“? (Figure 2.6.17).



davebishopcricket @bishopdave1965 · Oct 15, 2020

Replying to @Max17022629 and @richardlehnerdt

Agree with you wholeheartedly, at least the world media is starting to take notice. These EFFing people are shit stirrers and will stop at nothing. Just like Hitler, look what happened then.

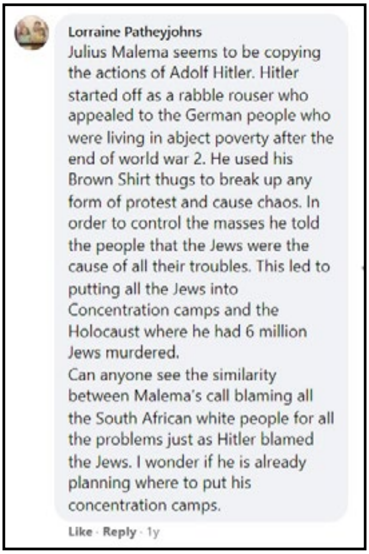
1 1

Max @Max17022629 · Oct 15, 2020

Yeah what happened then was genocide on the largest scale ever seen. We can't let that happen to us. We can wait for the word to react while they watch us being slaughtered and do nothing. We must fight back while they watching, then maybe they'll help.

1

Figure 2.6.16



Lorraine Patheyjohns

Julius Malema seems to be copying the actions of Adolf Hitler. Hitler started off as a rabble rouser who appealed to the German people who were living in abject poverty after the end of world war 2. He used his Brown Shirt thugs to break up any form of protest and cause chaos. In order to control the masses he told the people that the Jews were the cause of all their troubles. This led to putting all the Jews into Concentration camps and the Holocaust where he had 6 million Jews murdered.
Can anyone see the similarity between Malema's call blaming all the South African white people for all the problems just as Hitler blamed the Jews. I wonder if he is already planning where to put his concentration camps.

Figure 2.6.17

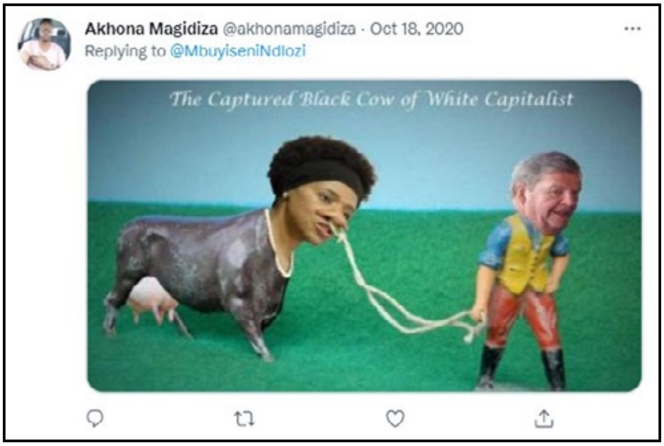
2.7 Fake Accounts, Foreigners, and Right-Left Convergence: Operation Dudula

One of the smears frequently directed at Thuli Madonsela was that she is in the pocket of "White Monopoly Capital (WMC)" (see Figure 2.7.1).[108] This image macro draws directly from the now defunct "WMCleaks" website, which rose to prominence early in 2017 as part of Bell Pottinger's social media strategy to deflect attention away from state capture.[109] To push this narrative, Bell Pottinger used platforms such as *The New Age* newspaper and Gupta-owned television channel ANN7, produced hate-filled articles on websites, and funded trolls and bots to spread its message on social media. Bell Pottinger helped create more than 100 fake Twitter profiles that retweeted content from accounts such as @economycapture and pushed hashtags such as #WhiteMonopolyCapital. The campaign produced 220 000 tweets.[110] The WMCleaks website and the accounts that amplified its content have since been shut down. Yet the refrain that Madonsela was "captured by WMC" continues to reverberate on social media. Unfortunately, as we will see below, others with malign intent have taken note of the power of fake accounts.

[108] Here Thuli Madonsela's face has been superimposed on to a cow that is being led by Johann Rupert.

[109] Following what were referred to as the #GuptaLeaks, the release of a trove of information that confirmed the process of state capture, the Guptas (through their company Oakbay Investments) hired Bell Pottinger to produce a public relations campaign that would divert attention away from them and on to their enemies. The head of this campaign wrote a letter to Jacob Zuma's son stating that "[t]he key to any political messaging is repetition and we will need to use every media channel that we can, to let our message take seed and to grow". This message became the narrative that whites in South Africa had seized state resources while they deprived blacks of education and jobs while also developing a campaign to besmirch a number of leading academics, journalists and business figures in South Africa. Ronel Rensburg, 'State Capture and the Demise of Bell Pottinger: Misusing Public Relations to Shape Future Kakistocracies?' in Krishnamurthy Sriramesh & Dejan Verčič (eds.), *The Global Public Relations Handbook: Theory, Research, and Practice* [3 ed.] (New York & London: Routledge, 2020), at page 85.

[110] Ronel Rensburg, 'State Capture and the Demise of Bell Pottinger: Misusing Public Relations to Shape Future Kakistocracies?' in Krishnamurthy Sriramesh & Dejan Verčič (eds.), *The Global Public Relations Handbook: Theory, Research, and Practice* [3 ed.] (New York & London: Routledge, 2020), at pages 90–91.



Akhona Magidiza @akhonamagidiza · Oct 18, 2020

Replying to @MbuyiseniNdlozi

Figure 2.7.1

Bot accounts (automated accounts that are created en masse to artificially amplify posts or topics) and 'sock puppet accounts' (fake accounts controlled by real people) have proliferated on social-media platforms globally. Facebook alone removed 15 billion fake accounts over the last two years, and it estimates that more than 90 million accounts (5% of its profiles) are fake.[111]

One such 'sock puppet account' on Twitter has had a significant impact on the development of the #PutSouthAfricafirst movement (which has since morphed into #PutSouthAfricansFirst). The movement demands that the government and the private sector privilege South Africans over foreign nationals and blames the latter for crime and other social issues. It has become increasingly influential on social media over the last two years; #PutSouthAfricansFirst was tenth on the list of all hashtags used in South Africa in 2020 (Figure 2.7.2).[112]

[111] Jack Nicas, 'Why can't the social networks stop fake accounts?', *The New York Times*, 8 December 2020. https://www.nytimes.com/2020/12/08/technology/why-cant-the-social-networks-stop-fake-accounts.html

[112] Centre for Analytics and Behavioural Change, 'Putting xenophobia first: Analysing the hashtags behind the Twitter campaigns', *Daily Maverick*, 20 Feb 2022. https://www.dailymaverick.co.za/article/2022-02-20-putting-xenophobia-first-analysing-the-hashtags-behind-the-twitter-campaigns/

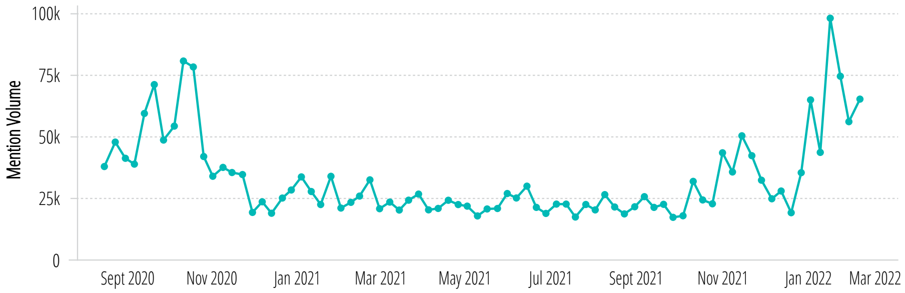


Figure 2.7.2: *The volume of the entire '#PutSouthAfricafirst' conversation (ie, associated hashtags) between 18 August 2020 and 13 February 2022 (weekly).*[113]

Analysis by Kyle Findlay reveals that the #PutSouthAfricafirst movement has two sides: a 'legitimate public side', represented by Herman Mashaba's Action SA party and the African Transformation Movement (a party with strong Radical Economic Transformation links),[114] and an 'anonymous side' that propagates attacks on social media on foreigners from influencers such as @uLerato_Pillay and @landback_.[115] @uLerato_Pillay was shown to have links with Mario Khumalo's South African First party,[116] an explicitly anti-immigrant party that has claimed that if it were elected, foreign nationals would have 48 hours to leave South Africa before the borders are sealed.[117]

A key offshoot of this movement has been a group that refers to itself as "Operation Dudula" that has advertised its marches and spread its message using social media (see figures 2.7.3 and 2.7.4).[118] 'Thato Moodley' (figures 2.7.5 and 2.7.6), whose handle was @uLeratoPillay1, was particularly active in promoting Operation

[113] https://www.talkwalker.com/blog/social-media-stats-south-africa; Centre for Analytics and Behavioural Change, 'Putting xenophobia first: Analysing the hashtags behind the Twitter campaigns', *Daily Maverick*, 20 Feb 2022. https://www.dailymaverick.co.za/article/2022-02-20-putting-xenophobia-first-analysing-the-hashtags-behind-the-twitter-campaigns/

[114] The Radical Economic Transformation (RET) is a faction of the ruling African Nationalist Congress that has led increasingly strident calls for more substantive redistribution of land and wealth in the country.

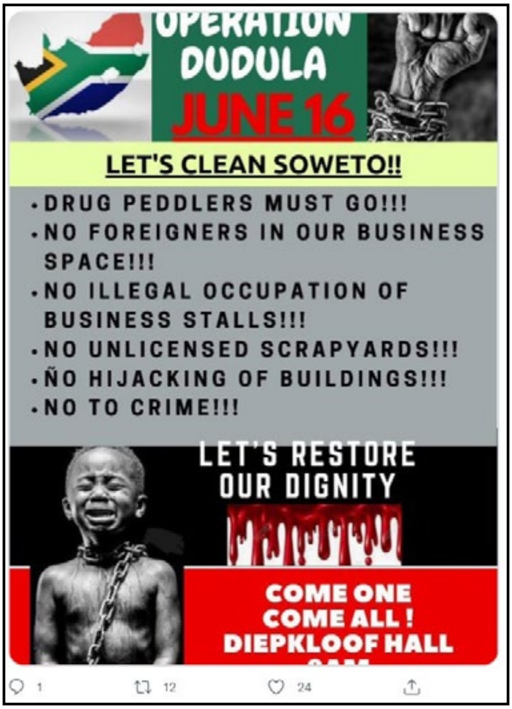
[115] Superlinear, 'Xenophobia, nationalism & populism: What's going on with #PutSouthAfricansFirst?', 14 August 2020. https://www.superlinear.co.za/xenophobia-nationalism-populism-whats-going-on-with-putsouthafricansfirst/

[116] @DFRLab, 'Afrophobic South African Twitter account connected to nationalist political party', 3 July 2020. https://medium.com/dfrlab/afrophobic-south-african-twitter-account-connected-to-nationalist-political-party-7e7205cc8987

[117] Sipho Mabena, 'African First party that aims to evict foreign nationals', *BusinessDay*, 21 February 2017. https://www.businesslive.co.za/bd/national/2017-02-21-i-am-not-xenophobic-says-founder-of-south-african-first-party-that-aims-to-evict-foreign-nationals/

[118] The date of the march indicated here was a significant one. 16 June is Youth Day, a public holiday celebrating the Soweto Uprising of 1976. Read in conjunction with the images it suggests that the youth are once again under threat (marked by the image of a crying child in chains), this time from "illegal foreign nationals".

Dudula; indeed the @uLeratoPillay1 account was dedicated purely to popularising the objectives of Dudula and #PutSouthAfricafirst.



OPERATION DUDULA JUNE 16

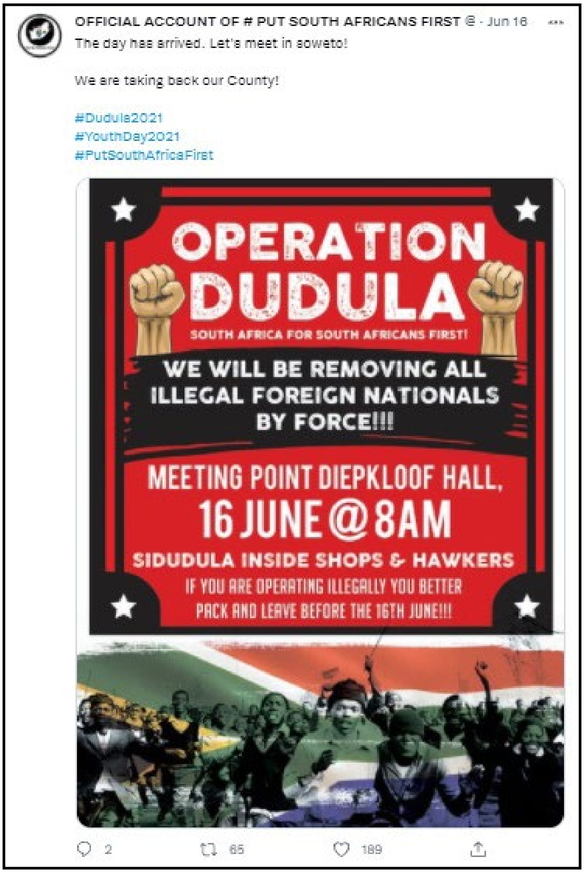
LET'S CLEAN SOWETO!!

- DRUG PEDDLERS MUST GO!!!
- NO FOREIGNERS IN OUR BUSINESS SPACE!!!
- NO ILLEGAL OCCUPATION OF BUSINESS STALLS!!!
- NO UNLICENSED SCRAPYARDS!!!
- ÑO HIJACKING OF BUILDINGS!!!
- NO TO CRIME!!!

LET'S RESTORE OUR DIGNITY

COME ONE COME ALL ! DIEPKLOOF HALL

Figure 2.7.3



OFFICIAL ACCOUNT OF # PUT SOUTH AFRICANS FIRST @ · Jun 16

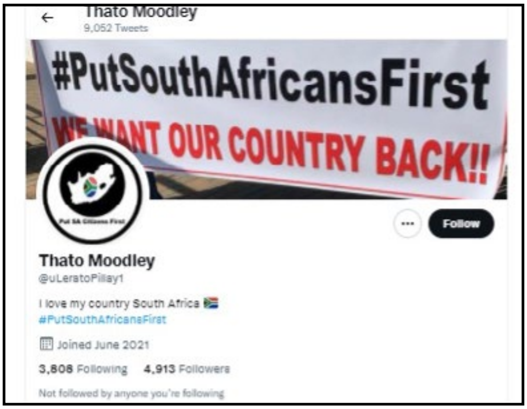
The day has arrived. Let's meet in soweto!

We are taking back our County!

#Dudula2021
#YouthDay2021
#PutSouthAfricaFirst

OPERATION
DUDULA
SOUTH AFRICA FOR SOUTH AFRICANS FIRST!
WE WILL BE REMOVING ALL ILLEGAL FOREIGN NATIONALS BY FORCE!!!
MEETING POINT DIEPKLOOF HALL,
16 JUNE @ 8AM
SIDUDULA INSIDE SHOPS & HAWKERS
IF YOU ARE OPERATING ILLEGALLY YOU BETTER PACK AND LEAVE BEFORE THE 16TH JUNE!!!
2
65
189

Figure 2.7.4



Thato Moodley
9,052 Tweets

Thato Moodley
@uLeratoPillay1

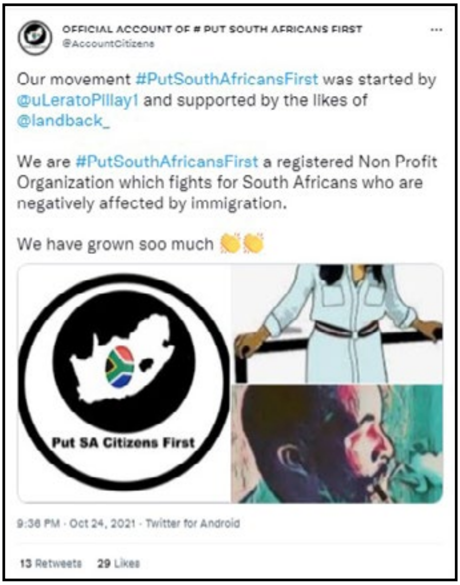
I love my country South Africa 🇿🇦
#PutSouthAfricansFirst

Joined June 2021

3,808 Following 4,913 Followers

Not followed by anyone you're following

Figure 2.7.5



OFFICIAL ACCOUNT OF # PUT SOUTH AFRICANS FIRST
@AccountCitizens

Our movement #PutSouthAfricansFirst was started by @uLeratoPillay1 and supported by the likes of @landback_

We are #PutSouthAfricansFirst a registered Non Profit Organization which fights for South Africans who are negatively affected by immigration.

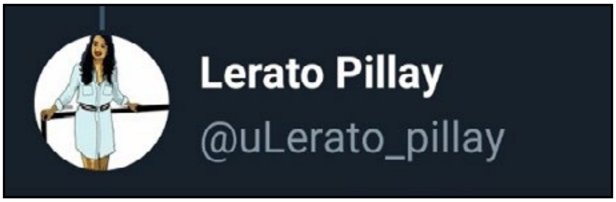
We have grown soo much 👏👏

9:36 PM · Oct 24, 2021 · Twitter for Android

13 Retweets 29 Likes

Figure 2.7.6

Thato Moodley's account drew on the memory, legacy, and ideology of *another* sock puppet account whose handle was @uLeratoPillay and operated under the name Lerato Pillay. That account deployed online disinformation targeting African foreigners and quickly developed a large following, reaching 45 million Twitter users before its suspension.[119] The Atlantic Council's Digital Forensic Research Lab exposed the owner of that account as Sifiso Gwala, a former member of the South African National Defence Force.[120] As seen in Figure 2.7.7, the name Lerato Pillay and profile image bore no relationship to the individual controlling the account.



Lerato Pillay
@uLerato_pillay

Figure 2.7.7

[119] Karen Allen, 'Digital vigilantism: Like fake news, has real-world consequences', *Daily Maverick*, 21 August 2020. https://www.dailymaverick.co.za/article/2020-08-21-digital-vigilantism-like-fake-news-has-real-world-consequences/.

[120] https://medium.com/dfrlab/afrophobic-south-african-twitter-account-connected-to-nationalist-political-party-7e7205cc8987

Thato Moodley seized upon the successes of the defunct @uLeratoPillay to promote Operation Dudula via the copycat account @uLeratoPillay1. While we have no idea who Moodley is, he (or she or they) simply made a minor change to the @uLeratoPillay handle. As can be seen from the network analysis in Figure 2.7.8, Moodley's copycat @uLeratoPillay1 account dominated our Dudula dataset by constantly posting and retweeting Afrophobic material. The account has now been suspended, months after it came into being. Other accounts in the dataset also paid homage to the "original" Lerato Pillay (see, for example, Figure 2.7.9).

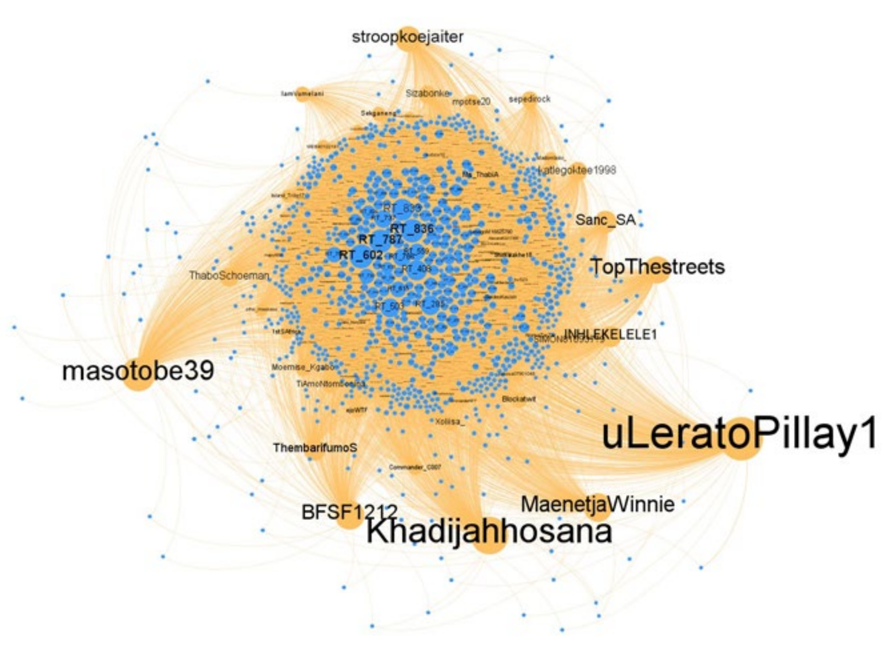
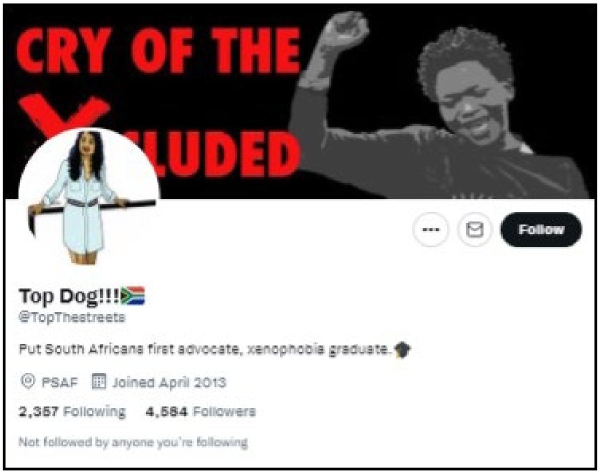


Figure 2.7.8



Top Dog!!!🇿🇦
@TopThestreets

Put South Africans first advocate, xenophobia graduate.🎓

PSAF Joined April 2013

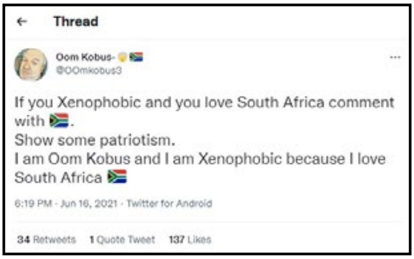
2,357 Following 4,584 Followers

Not followed by anyone you're following

Figure 2.7.9[121]

Despite "her" unmasking as Sifiso Gwala, Lerato Pillay maintained the status of a folk hero among supporters of Operation Dudula. Many users in our dataset incorporated the profile picture from the original Lerato Pillay account as part of their profiles, as well as paeans to her (and their own) "patriotism" in fighting for the survival of South Africa against unwanted immigrants. They drew an equivalence between being xenophobic and being patriotic: "I am Xenophobic because I love South Africa" (Figure 2.7.10).

Such references to patriotism were a striking feature of the content we examined. The popularity of the term may have reflected its resurgence in the United States during the Trump presidency. Figure 2.7.11 points to the potency and influence of imported words and concepts. Here, @landback_ has reworded MAGA as "Make Azania Great Again" and in the video incorporated into the tweet "WE CANT BREATH" is repurposed to agitate for the removal of immigrants. As the video continues, "WE CANT BREATH" is replaced by a list of countries from which immigrants have come.



Oom Kobus- 💡🇿🇦
@OOmkobus3

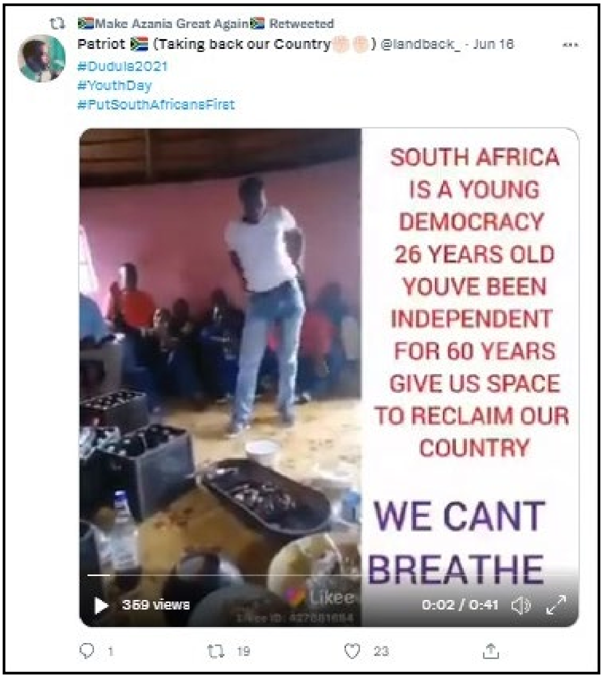
If you Xenophobic and you love South Africa comment with 🇿🇦.
Show some patriotism.
I am Oom Kobus and I am Xenophobic because I love South Africa 🇿🇦

6:19 PM · Jun 16, 2021 · Twitter for Android

34 Retweets 1 Quote Tweet 137 Likes

Figure 2.7.10

[121] Note how this incorporates Lerato Pillay's profile.



🇿🇦Make Azania Great Again🇿🇦 Retweeted

Patriot 🇿🇦 (Taking back our Country✊🏻✊🏻) @landback_ · Jun 16

#Dudula2021
#YouthDay
#PutSouthAfricansFirst

SOUTH AFRICA IS A YOUNG DEMOCRACY 26 YEARS OLD YOUVE BEEN INDEPENDENT FOR 60 YEARS GIVE US SPACE TO RECLAIM OUR COUNTRY

WE CANT BREATHE

Figure 2.7.11

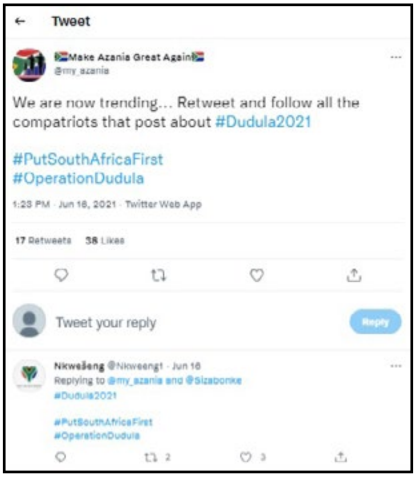
Thato Moodley's copycat account, much like that of the original @uLerato_Pillay, used derogatory terms profusely. For example, on 16 June 2021, Moodley posted a tweet stating, "Amakwerekwere. Amagrigamba. What other names do you know them as?".[122] In addition to spurring conversation, this deliberatively provocative question encouraged others to join them in breaching taboos, and Moodley seemed fully aware of the practical advantages of using vernacular derogatory terms to evade deletion or suspension by moderators.[123] Moodley and others were skilful Twitter users, employing follow-back campaigns to boost their number of followers and a wide range of hashtags to increase the visibility of their tweets in order to get Operation Dudula to trend.

[122] The first term, 'Amakwerekwere', according to Hashi Kenneth Tafira, refers to the phonetic sounds of African migrants that were incomprehensible to South Africans (though others suggest its origins may lie in a corruption of the word "korekore", the Korekore are a sub-group of the Shona people in neighbouring Zimbabwe). The term 'Amagrigamba', Tafira writes, refers to a person who came to South Africa with nothing but the clothes on their body. After a while, they return home wealthy, propertied and monied, all from the resources of the country. He notes that the term might be essentially economic but has now merged with racial identification. Hashi Kenneth Tafira, *Xenophobia in South Africa: A History* (Palgrave Macmillan, 2018), at page 24.

[123] When one user responded that Moodley should not use such names, Moodley's response was "When we call them the patriots English names our accounts get blocked. On my account I will refer to them as kweres".

They understood that Twitter's trending algorithm measures and rewards sharp spikes in hashtag use and were aware of the lurking threat of content moderators. On the morning of an Operation Dudula protest, for example, they were concerned that #OperationDudula #Dudula2021 was not trending. They suggested alternative hashtags (#YouthDay2021 and #sowetouprising) and blamed the problem on a conspiracy against them.

These fears were misplaced (Figure 2.7.12). A few hours later, @landback_ tweeted that Operation Dudula was a success: "Thank you ALL Patriotic South Africans who made it happen. Thanks to "xenophobic Twitter" 😂 for making it trend. You made everyone aware of the event 🤟🙏" (Figure 2.7.13). It is not clear if @landback_ was referring to users on Twitter who are xenophobic or those who denounced the movement, but given the number of reports at the time criticising the movement as xenophobic the latter is more likely. Here we probably see recognition that eliciting criticism can paradoxically serve as a mechanism to advance the profile of the movement by drawing attention on social media.



Tweet

🇿🇦Make Azania Great Again🇿🇦
@my_azania

We are now trending... Retweet and follow all the compatriots that post about #Dudula2021

#PutSouthAfricaFirst
#OperationDudula

1:23 PM · Jun 16, 2021 · Twitter Web App

17 Retweets 38 Likes

Tweet your reply

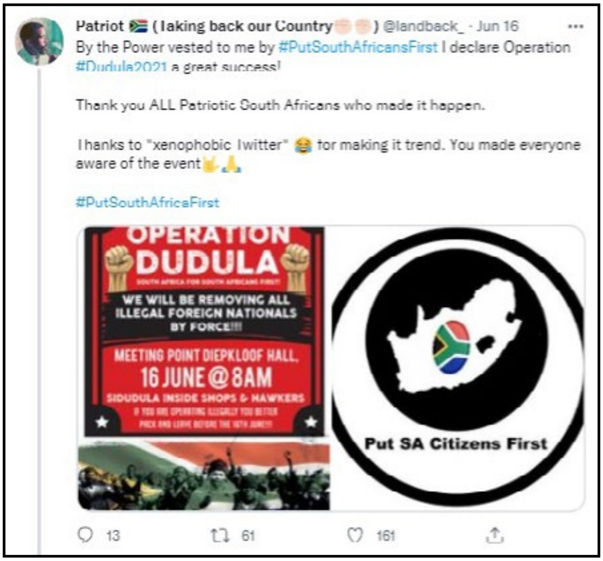
Reply

Nkwešeng @Nkweengt · Jun 16
Replying to @my_azania and @Sizabonke
#Dudula2021

#PutSouthAfricaFirst
#OperationDudula

2 3

Figure 2.7.12



Patriot 🇿🇦 (Taking back our Country✊🏻✊🏻) @landback_ · Jun 16

By the Power vested to me by #PutSouthAfricansFirst I declare Operation #Dudula2021 a great success!

Thank you ALL Patriotic South Africans who made it happen.

Thanks to "xenophobic Twitter" 😂 for making it trend. You made everyone aware of the event🤟🙏

#PutSouthAfricaFirst

OPERATION
DUDULA
SOUTH AFRICA FOR SOUTH AFRICANS FIRST!
WE WILL BE REMOVING ALL ILLEGAL FOREIGN NATIONALS BY FORCE!!!
MEETING POINT DIEPKLOOF HALL,
16 JUNE @ 8AM
SIDUDULA INSIDE SHOPS & HAWKERS
IF YOU ARE OPERATING ILLEGALLY YOU BETTER PACK AND LEAVE BEFORE THE 16TH JUNE!!!
Put SA Citizens First

13 61 161

Figure 2.7.13

In Figure 2.7.14 and elsewhere we see the embrace of the language and imagery of war. This was a constant across the dataset. Users wrote of 'no retreat' and 'no surrender', of fighting for South Africa and at times of killing immigrants (Figure 2.7.15). Often linked to this idea of fighting were claims of the need to purge the country of illegal immigrants. While the term 'purge' was used on occasion, more often 'clean' was preferred. Operation Dudula marches are often described as "clean-up operations", framing foreigners as dirt sullying the nation.

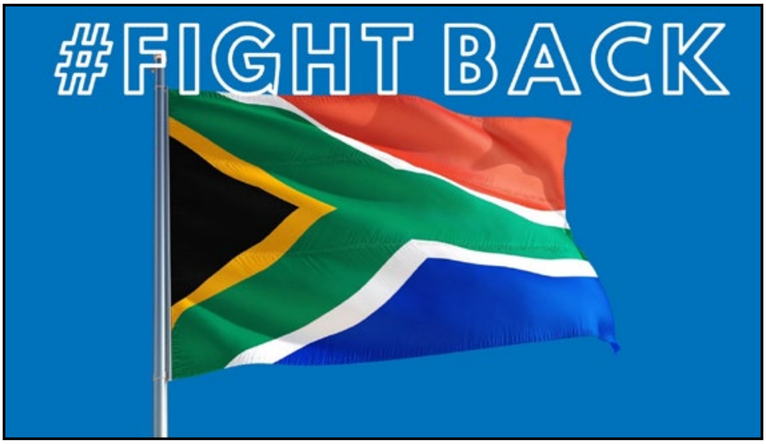
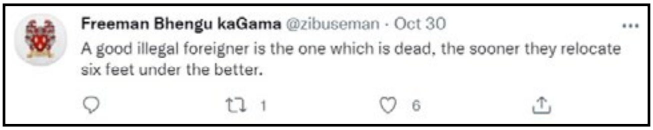


Figure 2.7.14



Freeman Bhengu kaGama @zibuseman · Oct 30

A good illegal foreigner is the one which is dead, the sooner they relocate six feet under the better.

Figure 2.7.15

The language of animalisation was equally popular. By implication, those who are animals do not warrant being treated as human beings. Such language was present across all the flashpoints we examined but were not as plentiful as we expected in relation to the Senekal and Brackenfell protests. As seen in some of the posts presented in this report, white users were referred to as “pigs“ or “pink pigs“, “insects” and on one occasion as “parasites“, while blacks users were referred to as “monkeys“, “baboons“, “dogs“, and on very rare occasions as “cockroaches“, “rats“ and “parasites“ (this last term was limited to Julius Malema and EFF protesters; see Figure 2.7.16).

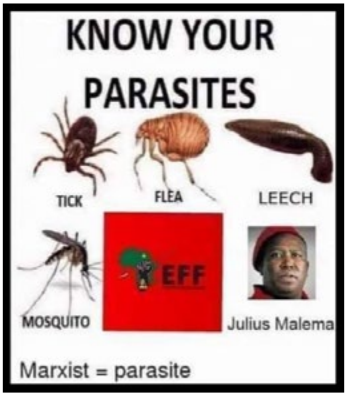
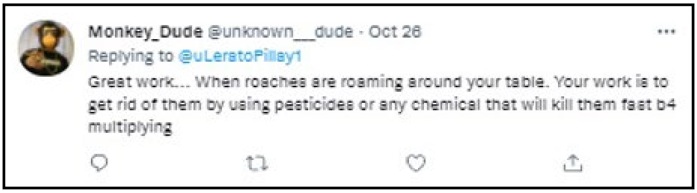


Figure 2.7.16

Yet animalising tropes took on a different character when it came to foreigners. They were sometimes referred to as “breeding like rats“ and described as cockroaches that needed to be eliminated (see Figure 2.7.17). Rats and cockroaches have a very particular stigma in the popular imagination of dirt and disease, and these ideas seemed to be encapsulated by one user who referred to foreigners as a “real pandemic in our Democracy“, a particularly evocative phrase during the Covid-19 pandemic.

The animalising tropes that were by far the most popular, however, were ones that referred to foreigners as “leeches“ and “parasites” (see, for example, figures 2.7.18 and 2.7.19). This is a particularly pernicious form of animalisation: by implication foreigners will drain the vitality of the body politic and even cause death, if not removed.

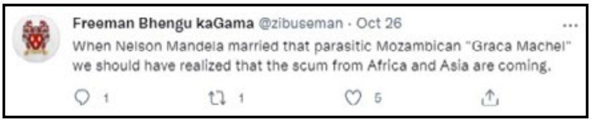


Monkey_Dude @unknown__dude · Oct 26

Replying to @uLeratoPillay1

Great work… When roaches are roaming around your table. Your work is to get rid of them by using pesticides or any chemical that will kill them fast b4 multiplying

Figure 2.7.17

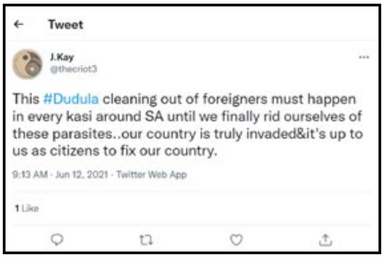


Freeman Bhengu kaGama @zibuseman · Oct 26

When Nelson Mandela married that parasitic Mozambican "Graca Machel" we should have realized that the scum from Africa and Asia are coming.

1 1 5

Figure 2.7.18



Tweet

J.Kay
@thecriot3

This #Dudula cleaning out of foreigners must happen in every kasi around SA until we finally rid ourselves of these parasites..our country is truly invaded&it's up to us as citizens to fix our country.

9:13 AM · Jun 12, 2021 · Twitter Web App

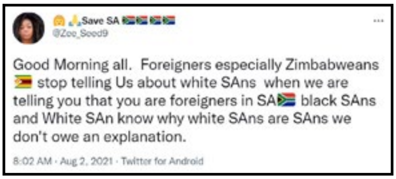
1 Like

Figure 2.7.19

The spectre of Lerato Pillay looms large in these calls to forcibly remove foreigners. As already described, this account had a substantial following. Unusually, it was embraced by both ostensibly white and black social media users. Curiously the fact that many thought Lerato Pillay was an Indian woman did not hinder "her" appeal. In this instance, appearing to be neither black nor white may have been an advantage as it allowed both the white right and black left to claim her.

Indeed, #OperationDudula was exceptional in the cases we examined for providing rare common ground and common cause for far-right white social media users as well as black social media users from the radical left. Here finally was something that they could agree on. Here there was the affirmation that both native-born blacks and whites belonged in South Africa. One black user pointed out that "We need to be careful of African immigrants and their stunts on white SAns. These people are here to devide us so our country can be a slum like other failed African states. Whites in this country belong here and not African immigrants". Another pointedly referred to Zimbabwe's failed land reform programme: "Black & white SAfricans know they need each other.U chased away white people from yo contries, took the land & OWN it. Instead of enjoying your land you'r following the same white people. Leave us & our whites alone, go back home & enjoy yo land in peace we'r enjoying landlessness". Such posts were often in response to claims that whites are settlers and therefore have even less right to remain in South Africa than African immigrants.

Figures 2.7.20 and 2.7.21 best encapsulate this rare interracial consensus and solidarity on South African social media between those on the radical left and the far right. The tragedy is that it is reached only through a shared hatred of African immigrants.

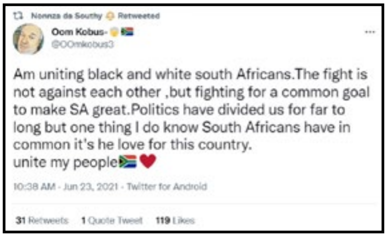


🙏Save SA 🇿🇦🇿🇦🇿🇦🇿🇦
@Zoe_Seed9

Good Morning all. Foreigners especially Zimbabweans 🇿🇼 stop telling Us about white SAns when we are telling you that you are foreigners in SA🇿🇦 black SAns and White SAn know why white SAns are SAns we don't owe an explanation.

8:02 AM · Aug 2, 2021 · Twitter for Android

Figure 2.7.20



Nonnza da Southy Retweeted
Oom Kobus- 🇿🇦
@OOmkobus3

Am uniting black and white south Africans.The fight is not against each other ,but fighting for a common goal to make SA great.Politics have divided us for far to long but one thing I do know South Africans have in common it's he love for this country.
unite my people🇿🇦❤️

10:38 AM · Jun 23, 2021 · Twitter for Android

31 Retweets 1 Quote Tweet 119 Likes

Figure 2.7.21

Operation Dudula demonstrates the potential for online communities to transition into real-world movements. Since its development as a small splinter movement of #PutSouthAfricansFirst, Operation Dudula has begun to formalise its structures and continues to grow under the leadership of Nhlanhla "Lux" Dlamini, a man who wears paramilitary-style camouflage gear at marches and speaks of "taking back" South Africa.[124] Its ambitions now also extend beyond Johannesburg. The group has opened branches in Cape Town and has indicated that it will soon do the same in Mpumalanga and Limpopo.[125] At the Cape Town launch, Dan Radebe said that they wanted to make the people of Cape Town aware of the "foreign invasion" in

[124] Thabo Myeni, 'What is Operation Dudula, South Africa's anti-migration vigilante?', *Al Jazeera*, 8 April 2022. https://www.aljazeera.com/features/2022/4/8/what-is-operation-dudula-s-africas-anti-immigration-vigilante (accessed 13 May 2022).

[125] Storm Simpson, 'Operation Dudula makes demands on Home Affairs, police at Cape Town launch', *The South African*, 16 May 2022. https://www.thesouthafrican.com/news/breaking-operation-dudula-makes-demands-on-home-affairs-police-at-cape-town-launch-patrick-mokgalusi-jonathan-buja-16-may-2022/ (accessed 16 May 2022).

the country.[126] More concerningly, Zandile Dubula revealed during the launch that Dudula was expanding its remit to target legal foreigners, stating, "even if you are legal in the country, if [you do not have] a scarce skill then you cannot be working. Now, it is illegal and legal immigrants. If you are legal here, you are not supposed to be working in a restaurant. Working in a restaurant doesn't require any special skill".[127]

When launched in April 2022 in Durban, participants sang anti-immigrant songs and often used the term 'amakwerekwere'.[128] Dan Radebe (the deputy chairperson of the movement) described Durban as a critical point as it "is the very same harbour they (illegal immigrants) are using as the point of entry for all the fake goods that have flooded our country, killing our textile industry which then affects the unemployment rate as well".[129]

This is simply another variation of the blaming of foreigners for South Africa's ills.[130] As a result, Operation Dudula, in both its online and real-world manifestations, has positioned itself as a legitimate voice for these grievances. It has the potential to morph into a movement with more political aims. Already political parties such as ActionSA and the Patriotic Alliance are working to co-opt the movement and its message.[131]

The potential impact of the violent rhetoric of the kind promoted by Operation Dudula was made tragically clear on 6 April 2022, when Mbodazwe Banajo "Elvis" Nyathi, a 43-year-old Zimbabwean, was beaten in front of his wife and burnt to death by a reported 30 people "who went around Diepsloot asking migrants to show their documents that permit them to be in South Africa legally".[132]

[126] Sibulele Kasa and Brandon Nel, 'Operation Dudula sets sights on Parklands', *IOL*, 15 May 2022. https://www.iol.co.za/weekend-argus/news/operation-dudula-sets-sights-on-parklands-be97f547-4820-409f-ae8d-97cc11548909 (accessed 15 May 2022).

[127] Philani Nombembe, 'Operation Dudula now targeting 'both legal and illegal immigrants', *TimesLIVE*, 15 May 2022. https://www.timeslive.co.za/news/south-africa/2022-05-15-operation-dudula-now-targeting-both-legal-and-illegal-immigrants/ (accessed 15 May 2022).

[128] Nokulunga Majola, 'Operation Dudula members march through Durban's city centre', *GroundUp*, 11 April 2022. https://www.groundup.org.za/article/operation-dudula-members-march-south-beach/ (accessed 13 May 2022).

[129] Rédaction Africanews, 'South Africa: Anti-immigration movement 'Operation Dudula' launched in Durban', *africanews*, 11 April 2022. https://www.africanews.com/2022/04/10/south-africa-anti-immigration-movement-operation-dudula-launched-in-durban// (accessed 13 May 2022)

[130] For a more detailed analysis of this process and how populist discourses such as this have become 'mainstreamed' see Johannes Machinya, 'Migration and Politics in South Africa Mainstreaming Anti-immigrant Populist Practice', *African Human Mobility Review*, Vol. 8, No 1 (Jan-Apr. 2022), pp. 59–78.

[131] Pearl Mncube, 'Operation Dudula: When deep-seated frustration meets prejudice and weak leadership', *Daily Maverick*, 20 April 2022. https://www.dailymaverick.co.za/opinionista/2022-04-20-operation-dudula-when-deep-seated-frustration-meets-prejudice-and-weak-leadership/ (accessed 14 May 2022).

[132] Naledi Sikhakhane, 'Mourners at Elvis Nyathi's memorial vent anger at Zimbabwe ambassador service', *Mail & Guardian,* 22 April 2022. https://mg.co.za/africa/2022-04-22-mourners-at-elvis-nyathis-memorial-vent-anger-at-zimbabwe-ambassador-service/#:~:text=Nyathi%20was%20beaten%20and%20burnt,his%20wife%20Nomsa%20Tshuma%2C%2038 (accessed 23 April 2022).

A report by the Centre for Analytics and Behavioural Change noted that online narratives that surfaced following Nyathi's death appear to have been manipulated to sow discord between South Africans and resident foreign nationals.[133] This tragic episode makes manifest the way in which social media in South Africa can create new communities and catalyse real-world actions, both good and ill. It thus behoves us to pay more attention to its potential to reshape the politics of South African society.

[133] Centre for Analytics and Behavioural Change, 'Posts about the death of Elvis Nyathi stoke the flames of xenophobia', *Daily Maverick*, 18 April 2022. https://www.dailymaverick.co.za/article/2022-04-18-posts-about-the-death-of-elvis-nyathi-stoke-the-flames-of-xenophobia/ (accessed 20 April 2022).

3 Content moderation

3.1 The Importance of Content Moderation

Content moderation (or perhaps the lack of it) has become one of the crucial questions of our time and shapes what we see on social-media platforms, which social-media platforms we join, what we post and how we post it.

When social-media platforms first developed, they were initially small and could survive by performing community-scale content moderation. In other words, those who set up a group would police its content with the help of some members. This process has proved untenable with the rapid rise of social-media platforms and the exponential increase in the amount of user-generated content being produced. The largest of these platforms developed in the US, where they were deemed to fall under the scope of Section 230 of the Communications Decency Act of 1996 (which was passed well before social media took on the size and status it has today). Section 230 states that "no provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider". These twenty-six words, Jeff Kosseff argues, created the modern internet.[134] It also gave the United States a competitive advantage when it came to the internet.[135] In addition, it allowed these platforms to practice a 'post first, remove later' policy, which has become a defining feature of social media.

Section 230 protected websites from lawsuits (with certain exceptions) if a user posted something illegal. In addition to this, the Communications Decency Act has a 'Good Samaritan' provision, which allows these platforms to "restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected." This allows platforms to develop their own community guidelines, to organise their content-moderation teams as they see fit (as long as they are removing copyright-protected and illegal content) and does not give platforms any guidance on what these community guidelines should be or how to enforce them.

As a result, the companies alone must develop and enforce these rules and decide how they will change as the scale of content moderation increases and as the

[134] Jeff Kosseff, *The Twenty-Six Words that Created the Internet* (Ithaca: Cornell University Press, 2019).

[135] Robyn Caplan, 'Content or Context Moderation: Artisanal, Community-Reliant, and Industrial Approaches' (Data & Society Research Institute, 2018), pp. 1–37, at page 27-28.

cultural environment shifts.[136] Section 230 and the Good Samaritan Provision gave technology companies leeway to set their own standards for content as well as giving them limited liability for most types of content posted by their users.[137] This became particularly important for social-media platforms whose whole business strategy is built on the constant publication of user-generated content at scale.

This law is coming under increasing pressure in the United States from both liberals and conservatives. Liberals feel that it allows social-media platforms to not remove enough content, while conservatives argue that it allows platforms to remove too much, in the process silencing conservative views.[138] Others have noted that the 'Good Samaritan' provision has merely shifted the public burden of regulation of speech onto platforms and they are doing this with few formal mechanisms for accountability or oversight.[139] Some have gone so far as to argue that it is now social-media companies, private companies who define what is blacklisted in their community standards, who have been given the outsourced project of defining our principles and morals regarding discussions in the public domain.[140] What is clear in both these views is that the question of content moderation has become increasingly central and even those who agree with the wide leeway provided by Section 230 are increasingly scrutinising the question of how platforms are making these decisions.[141]

As social-media platforms have grown, so has the problem of moderating them, which has posed both a logistics and a public relations problem for these platforms. Whereas in the early years of the internet, online communities only had to answer to their own users, the rapid increase in the quantity and variety of content, along with the fact that users are now often linked not only by bonds of community has meant that online harm now extends far beyond the platform on which it occurs.[142]

It is essential that readers are familiar with the content-moderation process in order to better understand this study: content moderation shaped what was left online for our annotators to examine. In order to understand the content of the archive, we examined social-media posts relating to the flashpoints that are the focus of this study – we need to first understand how the archive was curated.

This is more challenging than it seems. The content-moderation processes of social-media platforms are opaque, and each social-media platform undertakes the process in different ways. Section Three of this report, while not making any

[136] Ibid.

[137] Ibid., at page 4.

[138] GK Young, 'How much is too much: The difficulties of social media content moderation' in *Information & Communications Technology Law* (2021), pp. 1–16, at page 12.

[139] Robyn Caplan, 'Content or Context Moderation: Artisanal, Community-Reliant, and Industrial Approaches' (Data & Society Research Institute, 2018), pp. 1–37, at page 27.

[140] See, for example, Frederik Stjernfelt and Anne Mette Lauritzen, *Your Post Has Been Removed: Tech Giants and Freedom of Speech* (SpringerOpen, 2020).

[141] Robyn Caplan, 'Content or Context Moderation: Artisanal, Community-Reliant, and Industrial Approaches' (Data & Society Research Institute, 2018), pp. 1–37, at page 4.

[142] Tarleton Gillespie, 'Content moderation, AI, and the Question of Scale' in *Big Data and Society* (Jul.–Dec. 2020), pp. 1–5, at page 1.

claims to comprehensively unpacking the content-moderation practices of all social-media platforms, attempts to analyse what exactly commercial content moderation consists of, some of the key forms of content moderation undertaken by major social-media platforms, and what some of the crucial issues with these processes are.

3.2 What is Content Moderation?

It is important to point out from the start that content moderation on social-media platforms is no easy task. According to market and consumer data analytics firm Statista, Facebook had an estimated 2.85 billion active users sharing 4.75 billion items each day in 2020; Twitter had an estimated 314.9 million users posting approximately 500 million tweets per day.[143] Unsurprisingly then, even if only a fraction of these posts contain problematic material, content moderation on these platforms is an epic endeavour.

Given the large number of people using these and other social-media platforms, as well as the seeming ability of online discussion to mobilise and channel sentiment in the offline world, it is clear that hate speech on social-media platforms is one of the key issues of our day. As a result of the sheer number of users and posts, content moderation on these social-media platforms takes place on an industrial scale. Since users have come to expect that the content they post should appear instantaneously, a prepublication editorial review is impossible. As a result, detection of hate speech (and other forms of prohibited content) involves scouring what has already been posted and is already available for all to see.

Content moderation consists of the organised practice of screening user-generated content posted online in order to determine its appropriateness for a given site, locality, or jurisdiction.[144] It consists of both the detection of and interventions taken on content in conjunction with the rules imposed by platforms, the human labour and technologies used to screen content, as well as the mechanisms of adjudication, enforcement and appeal that support it.[145] Until the second half of

[143] See https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/ and https://www.statista.com/statistics/303681/twitter-users-worldwide/ (accessed on 6 August 2021)

[144] Ysabel Gerrard and Helen Thornham, 'Content Moderation Social Media's sexist assemblages' in *New Media & Society*, Vol. 22, No. 7 (2020), pp. 1266–1286, at page 1267.

[145] Ysabel Gerrard, 'The Best-Kept Secret in Tech' in Devan Rosen (ed.), *The Social Media Debate: Unpacking the Social, Psychological, and Cultural Effects of Social Media* (New York: Routledge, 2022), at page 77.

the 2010s, Ysabel Gerrard describes this process as “one of the tech world’s best-kept secrets“ with very few paying attention to the human labour that screened the internet for illicit content.[146] This has changed in the last five years with an increasing number of ethnographic studies of content moderation as well as other forms of academic and popular writing relating to the issue.[147]

Gerrard notes that the lack of attention to content moderation until recently can be ascribed to one of two things. First, a sign of good content moderation is its invisibility, “making it seem as though content just magically appears on a site, rather than there being some sort of curation process and a set of logics by which content is determined to be appropriate or inappropriate“.[148] It is important to note that all social-media companies undertake some form of content moderation even though their proficiencies may vary wildly. What content moderation consists of therefore very much depends on the platform in question and, often, its size.[149]

Early internet spaces worked on a ‘community reliant approach’. This consists of moderation done by members within the online community in question and is different to what Robyn Caplan labels ‘artisanal approaches’ (“where case-by-case governance is normally performed by between 5 and 200 workers“) and *industrial approaches* (“where tens of thousands of workers are employed to enforce rules made by a separate policy team“ – Facebook, for example, is estimated to have 15,000 content moderators).[150] These approaches should not be seen as mutually exclusive as companies often shift between these various approaches and may sometimes use multiple approaches concurrently. For example, Facebook’s content-moderation processes have changed as it has rapidly increased in size and scope, but until recently they still relied mainly on users as the first point of moderation.

The platforms that we are concerned with all use some form of industrial commercial content moderation, a process Caplan refers to as ‘the decision factory’. This industrial strategy depends on a highly formalised structure of organisation and rules where rule-making is typically separated, both geographically and

[146] Ibid.

[147] See for example the numerous articles relating to the issue of content moderation being published in popular tech sites such as *Wired* and *The Verge* as well as in old media such as *The Guardian* and *Wall Street Journal,* amongst others. For an academic engagement with the question of content moderation see Tarleton Gillespie, *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media* (New Haven & London: Yale University Press, 2018). For one of the more detailed ethnographies of human content moderation, see Sarah T Roberts, *Behind the Screen: Content Moderation in the Shadows of Social Media* (Newhaven: Yale University Press, 2019).

[148] Ysabel Gerrard, ‘The Best-Kept Secret in Tech’ in Devan Rosen (ed.), *The Social Media Debate: Unpacking the Social, Psychological, and Cultural Effects of Social Media* (New York: Routledge, 2022), at page 78.

[149] Ibid., at page 81.

[150] See Robyn Caplan, ‘Content or Context Moderation: Artisanal, Community-Reliant, and Industrial Approaches’ (Data & Society Research Institute, 2018), pp. 1–37 and Robyn Caplan, ‘The Artisan and the Decision Factory: The Organizational Dynamics of Private Speech Governance’ in Lucy Bernholz, Hélène Landemore & Rob Reich (eds.), *Digital Technology and Democratic Theory* (Chicago: University of Chicago Press, 2021), at pages 174–180.

organisationally from enforcement, “with policy teams distributed across the United States and Europe while those doing the enforcement of rules are located in places like the Philippines, Turkey, or India“.[151] A member of one of these policy teams described the process as trying to “create a ‘decision factory,’ which resembles more a ‘Toyota factory than it does a courtroom, in terms of actual moderation’”. The approach here, they went on to say, is to “take a complex thing, and break it into extremely small parts, so that you can routinize doing it over, and over, and over again“.[152] This formalisation and structure was required due to the need to ‘onboard’ moderators en masse and due to the high turnover of commercial content moderators.

Caplan therefore describes industrial content-moderation companies as “large-scale bureaucracies, with highly specialized teams, and distributions of responsibilities and powers. Interaction with platform users is largely confined to the system of flagging and review, done through a platform user interface (rather than contact with an employee)“.[153] As a result of the number of moderation decisions to be made, Caplan notes that such social-media platforms tend to collapse contexts in favour of establishing global rules, some of which make little sense in practice and often are unable to be sensitive to cultural concerns and context in the process of moderation.[154] It is therefore perhaps more useful for our purposes to use Sarah T Roberts’s term, *commercial content moderation*, which she describes as the monitoring and vetting of user-generated content for social-media platforms to ensure “compliance with legal and regulatory exigencies, site/ community guidelines, user agreements, and that it falls within norms of taste and acceptability for that site and its cultural context“.[155]

Differences in terms of the size, value, and missions of different social-media platforms will inevitably inform the approaches they take to content moderation.[156] However, despite various differences across platforms, each of the social-media platforms that form the focus of this study uses a combination of automated and human approaches. Tarleton Gillespie, a digital media specialist, has attempted to produce a general representation of the various forms of labour involved in the process of content moderation which can be seen in Figure 3.2.1. below.

Gillespie, however, goes on to point out that it is a mistake to think of platforms as filters when it comes to content moderation. Rather, he suggests that a better metaphor is to think of social-media platforms’ content-moderation practices as equivalent to trawling for fish:

[151] Robyn Caplan, ‘The Artisan and the Decision Factory: The Organizational Dynamics of Private Speech Governance’ in Lucy Bernholz, Hélène Landemore & Rob Reich (eds.), *Digital Technology and Democratic Theory* (Chicago: University of Chicago Press, 2021), at page 177.

[152] Ibid.

[153] Ibid.

[154] Ibid., at page 178.

[155] Sarah T Roberts, “Content Moderation,” in *Encyclopedia of Big Data* (New York: Springer, 2016).

[156] Robyn Caplan, ‘Content or Context Moderation: Artisanal, Community-Reliant, and Industrial Approaches’ (Data & Society Research Institute, 2018), pp. 1–37, at page 5.

> ***Platforms are filters only in the way that trawler fishing boats "filter" the ocean: they do not monitor what goes into the ocean, they can only sift through small parts at a time, and they cannot guarantee that they are catching everything, or that they aren't filtering out what should stay. This also means that even the most heinous content gets published, at least briefly, and the most criminal of behavior occurs and can have the impact it intended, before anything might be done in response. Content that violates site guidelines can remain for days, or years, in these wide oceans.***[157]

As seen in the analysis of coded datasets discussed in Section Two, tweets that cross the threshold for hate speech in the South African context remained on these platforms almost a year on from their date of original publication.

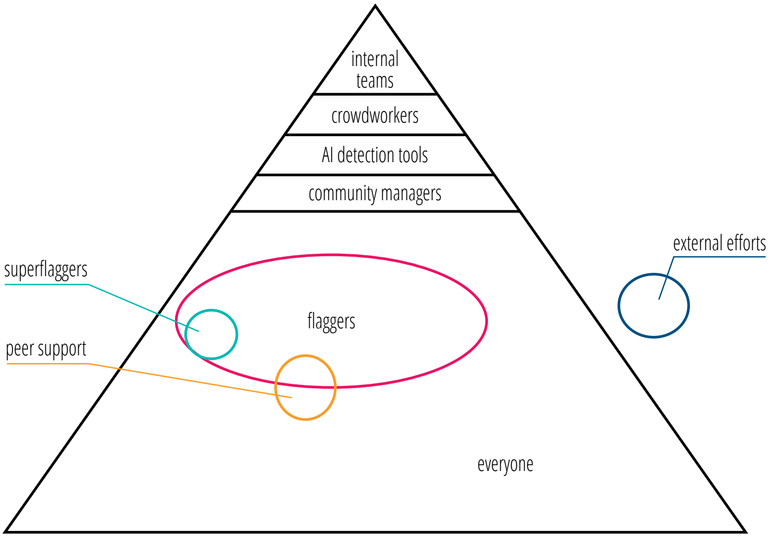


Figure 3.2.1: *The many forms of labour involved in platform moderation.*[158]

While we normally think of content moderation as simply relating to the moderation of content that appears on a particular user's account, it also consists of attempting to find and remove 'bot' and 'sock puppet' accounts. A key feature of the opacity of social-media platforms is the fact that there is no easy way for a lay person to know that a user is not whom they claim to be on their profile. While certain individuals may have their account verified for various reasons, for example

[157] Tarleton Gillespie, *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media* (New Haven & London: Yale University Press, 2018), at page 87.

[158] Image taken from Tarleton Gillespie, *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media* (New Haven & London: Yale University Press, 2018), at page 116.

influencers, this is the exception rather than the norm and even these verified accounts may be hacked by other parties. The use of anonymous or fake accounts on social media is therefore common. For example, the Carnegie Mellon University Centre for Informed Democracy and Social Cybersecurity reported in 2020 that the level of bot accounts (accounts that do not correspond to real people) across US and foreign elections, natural disasters, and other politicised events was normally between 10% and 20% and that this may have risen to between 45% and 60% when it came to Twitter accounts discussing Covid-19 in the US.[159] In addition to this there are also numerous 'sock puppet' accounts, fake accounts controlled by real people.

The scale of the issue can be seen by the fact that Facebook alone removed over 15 billion fake accounts over the last two years (see Figure 3.2.2). The vast majority of these were automated bot accounts that are created en masse to artificially amplify certain posts or topics, though the majority of these are caught fairly easily (as witnessed by a large number of removals). Jack Nicas, in a *New York Times* analysis of the issue shows that, while these tallies are impressive, those within the company note that the more telling metric is the prevalence of fake accounts, with Facebook estimating that 5% of its profiles are fake (more than 90 million accounts). This is due to social-media companies finding it more difficult to find and remove fake accounts that are created manually by a human. Nicas notes that these fakes are more pernicious because they look more believable and can be used to spread disinformation or to scam and defraud other users. Catching these accounts on Twitter is made even more complicated, as parody accounts are allowed (though these need to be clearly labelled).[160]

Nicas notes that the easiest way to combat this would be to require more documentation in order to create an account, something social-media companies are loath to do as it would make it more difficult for people to join their sites. The business models of these social-media platforms, Nicas goes on to note, are dependent on attracting users so they can sell more advertisements. For Twitter, this is also linked to its positioning as the 'free speech wing of the free speech party', as it prizes its users' anonymity in the belief that it enables dissidents to speak out against authoritarian governments.[161] As a result, in these cases (once again) the burden for flagging accounts often falls to users themselves.

[159] Karen Hao, 'Nearly half of Twitter accounts pushing to reopen America may be bots', *MIT Technology Review*, 21 May 2020. https://www.technologyreview.com/2020/05/21/1002105/covid-bot-twitter-accounts-push-to-reopen-america/

[160] Jack Nicas, 'Why can't the social networks stop fake accounts?', *The New York Times*, 8 December 2020. https://www.nytimes.com/2020/12/08/technology/why-cant-the-social-networks-stop-fake-accounts.html

[161] Ibid.

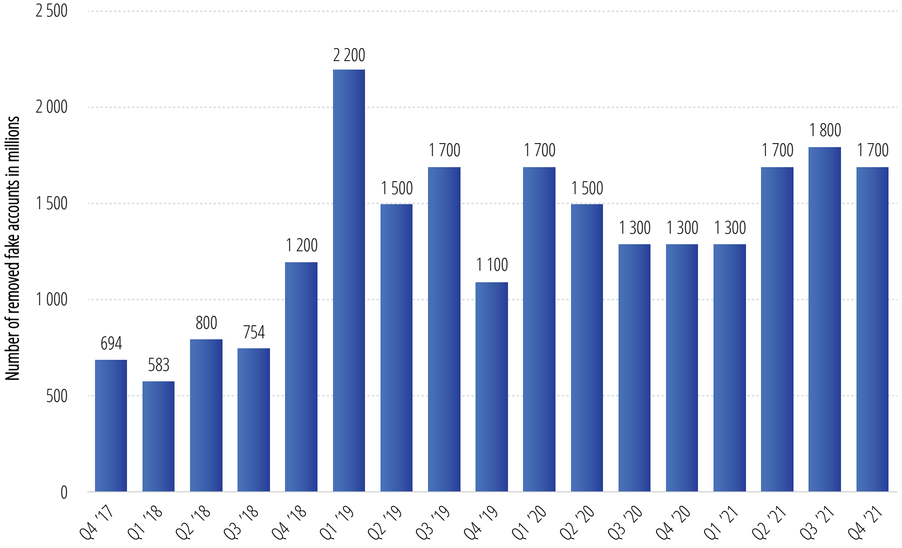


***Figure 3.2.2:** Global number of fake accounts taken action on by Facebook from the 4th quarter 2017 to the 4th quarter 2021.*[162]

3.3 The logic of opacity

The fact that people are now discussing content moderation (whether there is too much or too little of it) is a sign that its invisibility was not in fact a product of its effectiveness. Rather, Gerrard argues, its invisibility is a product of design. While social-media platforms have the power to shape what we do and do not see online, the public is often only aware of this process of content moderation when it affects them directly or when the process breaks down. Until recently, social-media companies have traditionally avoided acknowledging that content moderation takes place at all, who or what does whatever moderation work that does take place, what conditions this occurs under, and what their implementation policies are.[163]

[162] 'Global number of fake accounts taken action on by Facebook from 4th quarter 2017 to 4th quarter 2021', *Statista*. https://www.statista.com/statistics/1013474/facebook-fake-account-removal-quarter/ (accessed 23 January 2021).

[163] Ysabel Gerrard, 'The Best-Kept Secret in Tech' in Devan Rosen (ed.), *The Social Media Debate: Unpacking the Social, Psychological, and Cultural Effects of Social Media* (New York: Routledge, 2022), at page 78.

Sarah T Roberts has labelled this mysteriousness around the process of content moderation by social-media platforms "the logic of opacity". The aim of this opacity is to make platforms appear more objective in decision-making while also allowing users to assume that content moderation is done automatically and is driven by machine or machine-like rote behaviour that removes subjectivity. This protects these companies from questions around their policies and content-moderation processes.[164] Caplan notes that part of this opacity is due to the fact that "what constitutes content moderation, in terms of the practices, rules, and methods of enforcement, is still in flux"[165]. In what follows, we will attempt to make this process slightly less opaque and provide an overview of what commercial content-moderation work entails, and how it works on a human and technical level.

The first step of content moderation is the documents that underpin them. Facebook refers to these documents as 'community standards', Twitter as 'Rules and policies', and TikTok as 'community guidelines'. These all essentially refer to the same thing and in what follows we will be referring to them as 'community guidelines', a term that we feel better captures the actual role that they play. These documents, which are often not read at all by those who sign them, are what gives social-media platforms the power to remove content (or block its movement in various ways) no matter whether the content in question actually breaks local laws or the morals of the society that the individual poster belongs to.

3.4 Community guidelines relating to hate speech on Facebook, Twitter, and TikTok

Community guidelines are public-facing documents that are written in deliberately plain-spoken language that attempt to tell users how they are expected to behave and what kinds of content are acceptable and not acceptable.[166] Some of these rules are more stable than others, for example, rules against supporting terrorism, crime, sexual content involving minors etc. The three social-media platforms that this report focuses on, Facebook, Twitter, and TikTok, all also have rules against hate speech though they each use different terms to label these

[164] Sarah T Roberts, 'Digital detritus: 'Error' and the logic of opacity in social media content moderation ' in *First Monday*, Vol. 23, No. 3-5 (March, 2018).

[165] Robyn Caplan, 'The Artisan and the Decision Factory: The Organizational Dynamics of Private Speech Governance' in Lucy Bernholz, Hélène Landemore & Rob Reich (eds.), *Digital Technology and Democratic Theory* (Chicago: University of Chicago Press, 2021), at page 169.

[166] Ysabel Gerrard, 'The Best-Kept Secret in Tech' in Devan Rosen (ed.), *The Social Media Debate: Unpacking the Social, Psychological, and Cultural Effects of Social Media* (New York: Routledge, 2022), at page 80.

sections. It is referred to as 'hate speech' on Meta (Facebook's parent company), 'hateful conduct' on Twitter, and 'hateful behaviour' on TikTok. These guidelines and rules, however, are far less stable. This has often led to criticism of these policies for being too vague to govern speech for users all over the globe. It is also often unclear how exactly they are operationalised by content moderators or why particular sanctions are enforced (these sanctions may include banning or removal of content or accounts, demonetisation, de-ranking, or the inclusion of tags or warnings against problematic content).

Community guidelines relating to hate speech on Facebook

When content moderation began on Facebook in 2008, Facebook's guideline for moderators was only one page long, consisting of material to be removed, such as images of nudity and Hitler, while at the bottom of the page it simply stated: "Take down anything that makes you feel uncomfortable".[167] Since then, Facebook's guidelines have radically expanded but remained hidden from public view unless leaked, for example, when *The Guardian* published its 'Facebook Files' series in 2017, which consisted of exposés based on the internal manuals that were being used to moderate content.

When Max Fisher of *The New York Times* published an article on 'Facebook's Secret Rulebook for Global Political Speech', based on more than 1400 pages from these rulebooks, he argued that these extensive rules made the company "a far more powerful arbiter of global speech than has been publicly recognised or acknowledged by the company itself". The employee who leaked the document did so because of their fear that "the company was exercising too much power, with too little oversight — and making too many mistakes", allowing extremist language to flourish in some countries while censoring mainstream speech in others.[168] When it came to the issue of hate speech, the guidelines contained 200 "jargon-filled, head-spinning pages" that moderators had to go through.[169]

Facebook finally made its moderation guidelines available to the public in full in 2018 (albeit in a heavily edited form). This was the first time that the public was given direct insight into how the company policed content. Since its release, its policy on hate speech has gone through twenty-one versions since 25 May 2018.[170] Their current policy reads:

> ***We do not allow hate speech on Facebook because it creates an environment of intimidation and exclusion and in some cases may promote real-world violence.***

[167] Frederik Stjernfelt and Anne Mette Lauritzen, *Your Post Has Been Removed: Tech Giants and Freedom of Speech* (SpringerOpen, 2020), at page 128.

[168] Max Fisher, 'Inside Facebook's Secret Rulebook for Global Political Speech', *The New York Times*, 27 December 2018.

[169] Ibid.

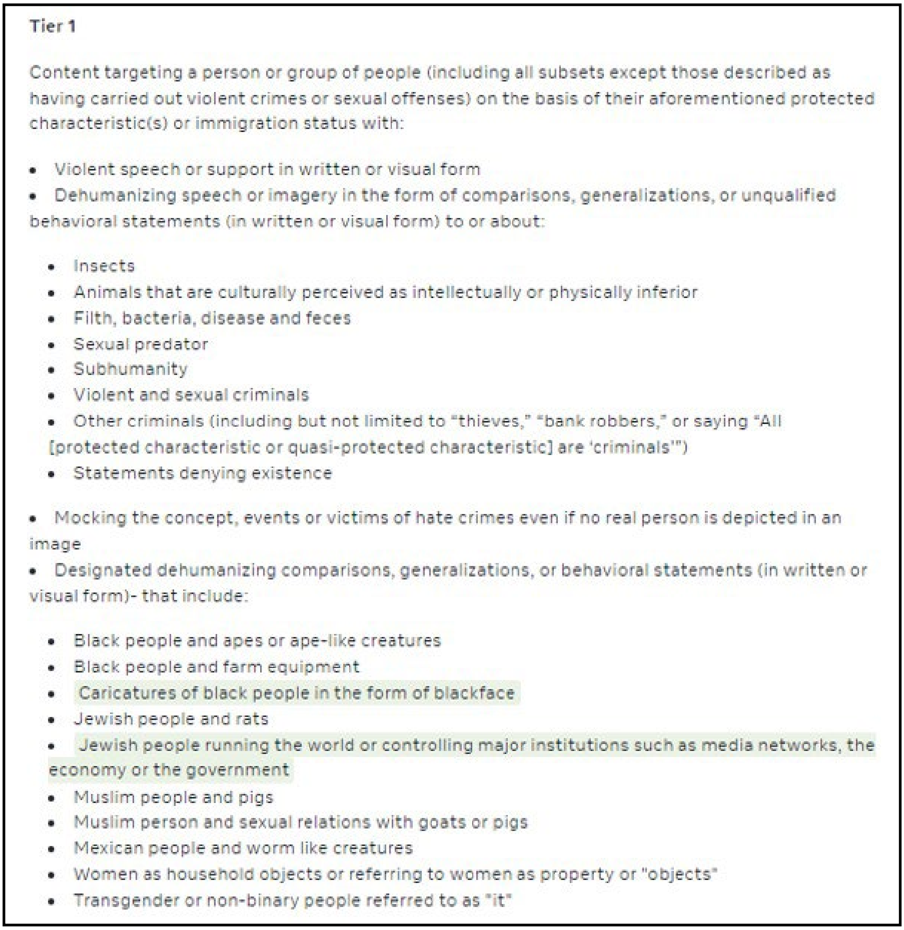
[170] The change log for these changes can be seen at https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/#policy-details

> ***We define hate speech as a direct attack on people based on what we call protected characteristics — race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability. We protect against attacks on the basis of age when age is paired with another protected characteristic, and also provide certain protections for immigration status. We define attack as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, or calls for exclusion or segregation.***[171]

This broad definition is then separated into three tiers of severity, each of which contains a long list of examples (see Figure 3.3). Facebook claims that this tiered approach has made its policies more nuanced and precise but many, such as Sarah Jeong of *The Verge*, argue it consists of a "convoluted set of rules" that contain "a series of vague pronouncements peppered with brief interludes of oddly specific breakdowns — [that] might make you feel sorry for the moderator who's trying to apply them. Every time a bizarrely detailed exception is tacked on, you can almost imagine the actual case scenario that prompted them to revise the guidelines."[172]

[171] https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/#policy-details

[172] Sarah Jeong, 'Turns out Facebook moderation sucks because its guidelines suck', *The Verge*, 24 April 2018.

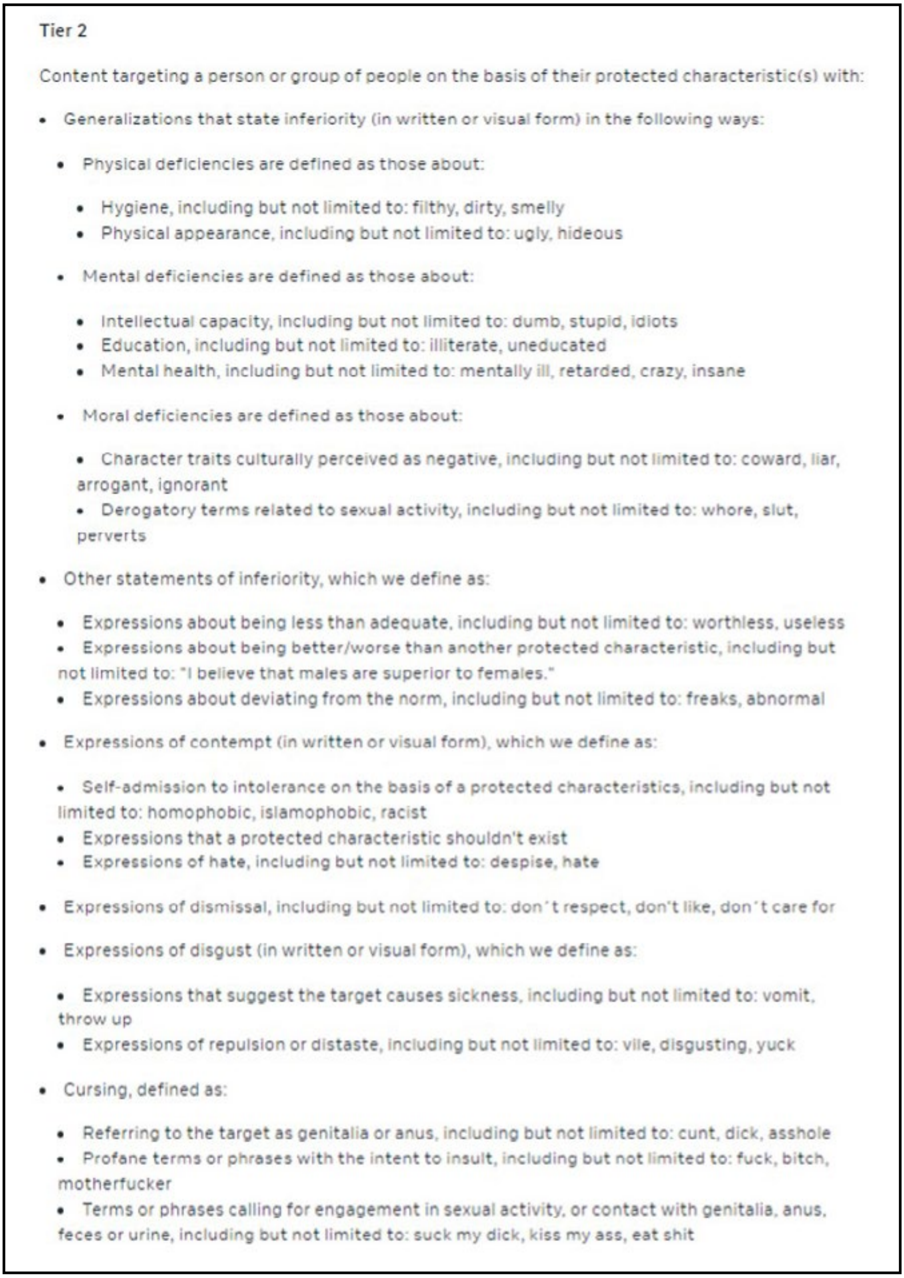


Tier 1

Content targeting a person or group of people (including all subsets except those described as having carried out violent crimes or sexual offenses) on the basis of their aforementioned protected characteristic(s) or immigration status with:

- Violent speech or support in written or visual form
- Dehumanizing speech or imagery in the form of comparisons, generalizations, or unqualified behavioral statements (in written or visual form) to or about:
  - Insects
  - Animals that are culturally perceived as intellectually or physically inferior
  - Filth, bacteria, disease and feces
  - Sexual predator
  - Subhumanity
  - Violent and sexual criminals
  - Other criminals (including but not limited to "thieves," "bank robbers," or saying "All [protected characteristic or quasi-protected characteristic] are 'criminals'")
  - Statements denying existence
- Mocking the concept, events or victims of hate crimes even if no real person is depicted in an image
- Designated dehumanizing comparisons, generalizations, or behavioral statements (in written or visual form)- that include:
  - Black people and apes or ape-like creatures
  - Black people and farm equipment
  - Caricatures of black people in the form of blackface
  - Jewish people and rats
  - Jewish people running the world or controlling major institutions such as media networks, the economy or the government
  - Muslim people and pigs
  - Muslim person and sexual relations with goats or pigs
  - Mexican people and worm like creatures
  - Women as household objects or referring to women as property or "objects"
  - Transgender or non-binary people referred to as "it"

***Figure 3.4.1:** The first of Facebook's three tiers of hate speech. Note that content highlighted in green refer to changes that have been made to the latest version of the guidelines.*

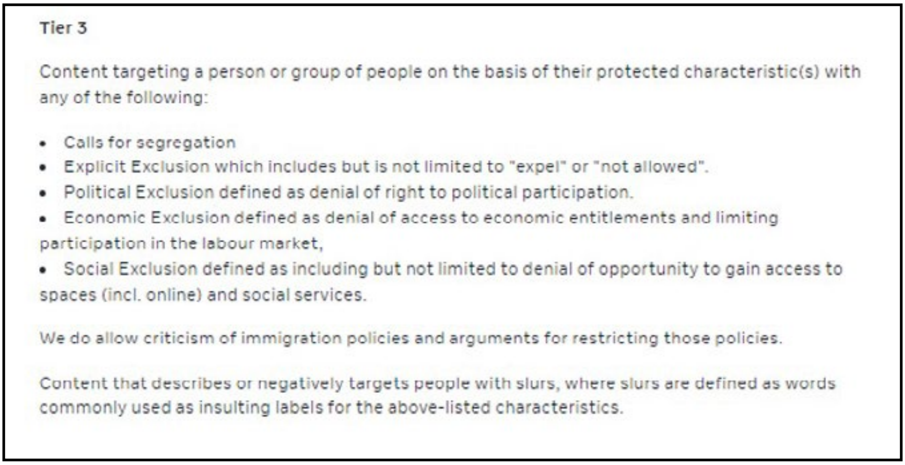


Tier 2

Content targeting a person or group of people on the basis of their protected characteristic(s) with:

- Generalizations that state inferiority (in written or visual form) in the following ways:
  - Physical deficiencies are defined as those about:
    - Hygiene, including but not limited to: filthy, dirty, smelly
    - Physical appearance, including but not limited to: ugly, hideous
  - Mental deficiencies are defined as those about:
    - Intellectual capacity, including but not limited to: dumb, stupid, idiots
    - Education, including but not limited to: illiterate, uneducated
    - Mental health, including but not limited to: mentally ill, retarded, crazy, insane
  - Moral deficiencies are defined as those about:
    - Character traits culturally perceived as negative, including but not limited to: coward, liar, arrogant, ignorant
    - Derogatory terms related to sexual activity, including but not limited to: whore, slut, perverts
- Other statements of inferiority, which we define as:
  - Expressions about being less than adequate, including but not limited to: worthless, useless
  - Expressions about being better/worse than another protected characteristic, including but not limited to: "I believe that males are superior to females."
  - Expressions about deviating from the norm, including but not limited to: freaks, abnormal
- Expressions of contempt (in written or visual form), which we define as:
  - Self-admission to intolerance on the basis of a protected characteristics, including but not limited to: homophobic, islamophobic, racist
  - Expressions that a protected characteristic shouldn't exist
  - Expressions of hate, including but not limited to: despise, hate
- Expressions of dismissal, including but not limited to: don´t respect, don't like, don´t care for
- Expressions of disgust (in written or visual form), which we define as:
  - Expressions that suggest the target causes sickness, including but not limited to: vomit, throw up
  - Expressions of repulsion or distaste, including but not limited to: vile, disgusting, yuck
- Cursing, defined as:
  - Referring to the target as genitalia or anus, including but not limited to: cunt, dick, asshole
  - Profane terms or phrases with the intent to insult, including but not limited to: fuck, bitch, motherfucker
  - Terms or phrases calling for engagement in sexual activity, or contact with genitalia, anus, feces or urine, including but not limited to: suck my dick, kiss my ass, eat shit

Figure 3.4.2: *The second of Facebook's three tiers of hate speech.*



Tier 3

Content targeting a person or group of people on the basis of their protected characteristic(s) with any of the following:

- Calls for segregation
- Explicit Exclusion which includes but is not limited to "expel" or "not allowed".
- Political Exclusion defined as denial of right to political participation.
- Economic Exclusion defined as denial of access to economic entitlements and limiting participation in the labour market,
- Social Exclusion defined as including but not limited to denial of opportunity to gain access to spaces (incl. online) and social services.

We do allow criticism of immigration policies and arguments for restricting those policies.

Content that describes or negatively targets people with slurs, where slurs are defined as words commonly used as insulting labels for the above-listed characteristics.

Figure 3.4.3: *The third of Facebook's three tiers of hate speech.*

There are a few points worth making here. These guidelines were only released after a great deal of effort on the part of various civic society groups and only at great risk to those who leaked the documents as Facebook and the company's commercial content moderation has been outsourced to make these employees sign strict Non-Disclosure Agreements. These Non-Disclosure Agreements are often heavy-handed attempts to keep these guidelines hidden from the broader public, which is unsurprising given the contradictions that riddle them. These hard-won guidelines may now provide us with an insight into the principles underpinning Facebook's policy on content removal, but they do not actually state how these policies relate to the enforcement procedure.

When it comes to taking action, Facebook describes a three-part approach. They may remove the material once they become aware of it, they may reduce the distribution of such content even if it does not meet the standard for removal under their policies or they may add a warning to potentially sensitive or misleading content. Depending on the number of 'strikes' that an account has, an account that falls foul of these guidelines may be restricted (where from the second strike onwards, you are restricted from creating and posting content for a set number of days leading to a 30-day restriction when an account has five or more strikes) or disabled. The disabling of an account can occur at any time, depending on the seriousness of the infraction.

This process of moderating hate speech on Facebook is in many ways a product of the commercial underpinnings of this particular social-media platform. Facebook developed to provide connections across users, but it very rapidly focused on ways to monetise these connections and their users. Facebook's profits (as is the case with most social-media platforms) are generated through advertising. Purchasing advertising on Facebook allows you to target users based on their location, demographic, and profile information and the advertisements will appear in your target group's sidebar or in their newsfeed. A 2017 *ProPublica* report highlighted how, unlike traditional media companies that select the audiences they offer advertisers, Facebook generates its advertising categories based on what users explicitly share with Facebook and what they implicitly convey through their online activity.[173]

The report went on to show how Facebook enabled advertisers to direct their advertisements to those who expressed interest in the topics of 'Jew hater'. When *ProPublica* widened the categories to include 'German Schutzstaffel' (the Nazi SS), 'Nazi Party', and 'National Democratic Party' (a far-right ultranationalist political party in Germany), the number had increased to 194 600 potential viewers. The advert was approved with the only change being the replacement of the ad category 'Jew hater' with 'Antysemityzm' (the Polish word for antisemitism). The results of the campaign were sent to *ProPublica* a few days later and the three advertisements they posted reached 5897 people, generated 101 clicks and 13 'engagements' (which could be a 'like', a 'share' or a comment on the post). A report a year later by *The Intercept* showed that Facebook was selling advertisers the ability to market to those with an interest in the 'white genocide' myth as "white genocide conspiracy theory" was a pre-defined 'detailed targeting' criterion (consisting of 168 000 users who were defined as "people who have expressed an interest or like pages related to White genocide conspiracy theory"). Other suggested advertising targets included mentions of South Africa where, as we saw in Section Two, the white genocide myth is a common trope.[174]

Key to Facebook's model is to have as many users as possible producing and consuming content that makes them remain online for as long as possible and produce as many data points for targeted advertising as possible. Facebook has aggressively expanded since its inception and has a pattern of behaviour that suggests one of its key aims has become to monetise data produced by its users while on the platform. This model requires Facebook to keep expanding to increase profits and to keep as many users as possible while also protecting the brand and advertisers from being associated with problematic content. As a result, Facebook tends to have a more conservative approach to content moderation and, as we will see below, it has been the most aggressive in producing automated content-moderation systems to protect its brand while also producing tiered layers of punishment to encourage users to remain on the platform while moderating their behaviour.

Community guidelines relating to hateful conduct on Twitter

Unlike Facebook, Twitter's hate speech guidelines are narrower, with more precise definitions provided, but the current guidelines have also changed over time. While its 2009-2015 preamble ended with the promise not to "censor user content, except in limited circumstances", a 2015 update to the preamble spoke about

[173] Julia Angwin, Madeleine Varner and Ariana Tobin, 'Machine Bias: Facebook Enabled Advertisers to Reach 'Jew Haters'', *ProPublica*, 14 September 2017.

[174] Sam Biddle, 'Facebook allowed advertisers to target users interested in "white genocide" – even in the wake of Pittsburgh massacre', *The Intercept*, 2 November 2018.

finding a balance between sharing content and "protect[ing] the experience and safety of people who use Twitter". This occurred soon after Twitter streamlined the reporting of abuse by allowing users to flag it more easily and streamlining the blocking of users. In 2017, new rules were introduced to curb abusive and threatening content in usernames and profiles, and between 2018 and 2020 their rules against hateful conduct and dehumanising speech were fleshed out.[175]

Its guidelines for hateful content as of 28 January 2022 are as follows:

> ***Hateful conduct: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories.***
>
> ***Hateful imagery and display names: You may not use hateful images or symbols in your profile image or profile header. You also may not use your username, display name, or profile bio to engage in abusive behavior, such as targeted harassment or expressing hate towards a person, group, or protected category.***[176]

This is followed by a section detailing when this will apply, along with some broad examples which consist of the following sections:

- Violent threats;
- Wishing, hoping or calling for serious harm on a person or group of people;
- References to mass murder, violent events, or specific means of violence where protected groups have been the primary targets or victims;
- Incitement against protected categories;
- Repeated and/or non-consensual slurs, epithets, racist and sexist tropes, or other content that degrades someone;
- Hateful imagery.

It is interesting to note that in most cases the sanction of tweets defined as 'violent threats' will lead to immediate and permanent suspension of an account. 'Repeated and/or non-consensual slurs, epithets, racist and sexist tropes, or other content that degrades someone', in "severe [cases] where the primary intent is to harass or intimidate others, may require Tweet removal" while 'moderate, isolated usage' may lead to the limitation of tweet visibility. For all other offences, Twitter's rules and policies page notes that the following potential consequences may be applied:

[175] Daniel Konikoff, 'Gatekeepers of toxicity: Reconceptualizing Twitter's abuse and hate speech policies in *Policy and Internet* (2021), pp. 502–521, at page 505.

[176] https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

- Down-ranking tweets in replies, except when the user follows the Tweet author.
- Making tweets ineligible for amplification in Top search results and/or on timelines for users who don't follow the tweet author.
- Excluding tweets and/or accounts in email or in-product recommendations.
- Requiring tweet removal.
  For example, we may ask someone to remove the violating content and serve a period of time in read-only mode before they can tweet again. Subsequent violations will lead to longer read-only periods and may eventually result in permanent suspension.
- Suspending accounts whose primary use we've determined is to engage in hateful conduct as defined in this policy, or who have shared violent threats.[177]

Twitter's policies on hate and abuse have grown more refined and precise over time while also ballooning in scope (the company's original rules were only 568 words in total for all possible infractions). The incremental changes, Sarah Jeong suggests, were introduced in response to legal threats or big news stories and suggest Twitter's repositioning from an 'anti-censorship platform' which once referred to itself as "the free speech wing of the free speech party" to the pragmatic reality of running a for-profit business, with rules changing whenever something threatened its bottom line (mainly made up of advertising revenue).[178]

It has, however, attempted to maintain its mythos as an anti-censorship platform. For example, the company claimed, following its addition of a raft of new changes, that their aim was to "[do] a better job combating abuse without chilling or silencing speech". When challenged that these changes effectively banned hate speech, the company responded by claiming that the company does not prohibit hate speech but 'hateful conduct', which it claimed differs from hate speech as the latter focuses on words, while they were prohibiting incitement to violence. Offensive and controversial viewpoints were still permitted. These shifts suggest the tensions in Twitter's view of itself as ideological protectors of free speech and moving more closely to what Jeong refers to as "the ideologically-unburdened censoriousness of Facebook and Instagram".[179]

Konikoff argues that Twitter attempts to weave "freedom rhetoric into almost every policy page it hosts, is forthright in touting an individualist ethos, and espouses traditional democratic ideals of equality, participation, and liberty".[180] This is also true of its 'hateful conduct' policy, which only bars hateful conduct to protect the free speech of others by arguing that threatening behaviour affects these users' safety or comfort on the platform. This is less broad than Facebook's

[177] Ibid.

[178] Sarah Jeong, 'The History of Twitter's Rules', Vice, 14 January 2016. https://www.vice.com/en/article/z43xw3/the-history-of-twitters-rules (accessed 26 January 2022).

[179] Ibid.

[180] Daniel Konikoff, 'Gatekeepers of toxicity: Reconceptualizing Twitter s abuse and hate speech policies in *Policy and Internet* (2021), pp. 502–521, at page 511.

policy which bars hate speech on more extensive grounds: “because it creates an environment of intimidation and exclusion and in some cases may promote real-world violence“.[181] This ‘free-speech-over-everything’ philosophy that encourages as few barriers as possible, Konikoff notes, permits users to generate toxic content without encouraging them to “withhold“ their own posts when they could potentially cause harm.[182]

When users fall foul of Twitter’s policies, there are three categories of sanction: tweet-level enforcement, direct-message level enforcement, and account-level enforcement. The punishment in each of these levels also varies according to the perceived seriousness of the offence. For example, Twitter can put a flagged account in read-only mode until a review is conducted of the profile; at the direct-message level, Twitter can temporarily limit a user’s ability to contact other users. At the tweet-level, Twitter may limit a tweet’s visibility by making it less visible in search results, replies, and on users’ timelines.[183] However, Twitter explicitly states that it initially focuses on the tweet level to ensure it is not being overly harsh with an otherwise ‘healthy’ account and it is explicit regarding its reluctance to enforce rules too sternly.

In addition to starting from a position that users do not intend to violate their rules, Twitter also delegates the task of gatekeeping to its users. While most social-media platforms now encourage their users to play a crucial part in gatekeeping on the sites by flagging questionable content to be sent to content moderators for review, Konikoff notes that Twitter also devolves responsibility onto users as the primary gatekeepers of hateful and abusive content by advising them to “unfollow“ other users who post material they do not like and, if the abusive behaviour continues, to then block the account. There then follows several caveats urging the user to reconsider whether the content actually qualifies as online abuse and whether users can resolve their differences, before then suggesting that they consider reporting the behaviour. Reporting is thus framed as an exceptional final step.

As is the case with all the social-media platforms under scrutiny here, there is no organisational gatekeeping, with material only being removed *after* being posted, and users who witness the material have to act as gatekeepers, although they are being increasingly joined by automated methods. Once material is flagged, the material is then sent to content moderators who make decisions on this material.[184] While it could be argued that this process of deputising users to flag content makes such platforms more democratic, there is much to be criticised about a process that allows hateful content to be seen by potentially a multitude of users before being removed.

Konikoff argues that, when taken together, this is part of a broader strategy by Twitter to absolve itself of enforcing its hate and abuse policies by transferring the “duty to enforce“ onto its users, who need to act as gatekeepers, meaning

[181] Ibid., at page 512.

[182] Ibid.

[183] Ibid.

[184] Ibid., at page 516.

Twitter's policies are in fact only effective if users actively flag hateful conduct. As Konikoff puts it, "[t]his filters Twitter's policies through individual user perceptions and forces the 'gated' users to be proactive gatekeepers of online hate and abuse if they want Twitter to do something about it".[185]

Even if users do diligently flag problematic content despite the repeated exhortations not to, there is little to suggest that Twitter has the capacity to enforce its own rules consistently, and in fact it often explicitly does not do so. Despite changes to its hateful conduct policies, there have been very few changes to Twitter's tangible enforcement mechanisms, as well as increasing critiques for the geographical inconsistency in its rule enforcement. For example, the de-platforming of Trump was compared by many activists to the lack of action against politicians in Sri Lanka, Myanmar, India, and Ethiopia.[186] Twitter also, by its own admission, unevenly applies its rules depending on the social status of the tweeter (a practice also carried out by other social-media platforms). For example, Twitter may only apply notices to tweets and prevent algorithmic elevation if the problematic tweet comes from an account by elected officials or those running for public office, have more than 100 000 followers, and have a verified account. Therefore, tweets by those with the largest followings may be exempt from the same forms of content moderation as the public.[187]

In addition to the above, as with other social-media platforms, there is no guarantee that the material will be removed following moderation (the process is neither consistent nor swift) and if it is removed, the users in question can contest this. While this is an important step to ensure that wrongly flagged accounts are not silenced, Konikoff suggests that all of these policies taken together highlight the reluctance to moderate content that is baked into Twitter's enforcement policies and he argues that this, along with the splitting of the gatekeeping role between the platform and its users, perpetuates abuse and hateful conduct.[188]

The content-moderation strategies of Twitter described above may, however, radically shift with the potential takeover of the social-media platform by Elon Musk. A long-time user of Twitter, Musk (the CEO of Tesla and SpaceX, who was declared by Forbes as the richest man in the world in 2022, with a net worth of $219 billion) bought a 9.2% share of the company on 4 April 2022, the first step in an attempted hostile takeover. On 14 April, Musk made a $43.4 billion offer to own the company outright, which would make it a private company. At the time of publication, it was unclear if this deal would go through as Musk has stated that the

[185] Ibid., at pages 513–514.

[186] Ibid., at page 506.

[187] Ysabel Gerrard, 'The Best-Kept Secret in Tech' in Devan Rosen (ed.), *The Social Media Debate: Unpacking the Social, Psychological, and Cultural Effects of Social Media* (New York: Routledge, 2022), at page 81. This policy can be seen at 'Defining public interest on Twitter', https://blog.twitter.com/en_us/topics/company/2019/publicinterest.

[188] Daniel Konikoff, 'Gatekeepers of toxicity: Reconceptualizing Twitter's abuse and hate speech policies in *Policy and Internet* (2021), pp. 502–521, at page 502 and 513.

acquisition was on hold due to fears about the level of automated 'bot' accounts that were present on the platform.[189]

Before this postponement, Musk, who has described himself as a "free speech absolutist", tweeted the following statement announcing the deal on 25 April:

> ***Free speech is the bedrock of a functioning democracy, and Twitter is the digital town square where matters vital to the future of humanity are debated. I also want to make Twitter better than ever by enhancing the product with new features, making the algorithms open source to increase trust, defeating the spam bots, and authenticating all humans.***

This has led to numerous criticisms, many of which boil down to the idea that providing equal and unmoderated access to all will in fact result in deeply iniquitous results, particularly for those who have already experienced historical harm.[190] As Michael Kleinman, the director of Amnesty International's Silicon Valley Initiative and an expert on Twitter harassment put it, "The more that people are harassed, the less likely they are to speak out […] What I fear is the voices that we most need to hear, the voices most impacted by structural inequalities or racism, it's those voices that will be silenced".[191]

Just as concerningly, many of Musk's answers to questions following the proposed takeover suggest a distinct lack of familiarity with the content-moderation process or its relationship to free speech laws and contestation around these laws.[192] One Twitter employee anonymously told *Time Magazine* that, while Musk's goals and Twitter's may be aligned, "the idea of bringing more free speech to the platform exposes his naiveté with respect to the nuts and bolts of content moderation" and

[189] Richard Waters, Hannah Murphy and Patrick McGee, 'Elon Musk raises prospect of lower price for Twitter deal', *Financial Times*, 17 May 2022. https://www.ft.com/content/2baac8e5-48ce-4aa3-a908-0ee2d5693c84 (accessed 17 May 2022).

[190] See, for example: Jessica Maddox, 'Elon Musk's comments about Twitter don't square with the social media platform's reality', *The Conversation*, 3 May 2022. https://theconversation.com/elon-musks-comments-about-twitter-dont-square-with-the-social-media-platforms-reality-182023 (accessed 13 May 2022); Adi Robertson, 'Elon Musk's Twitter Plans are a Huge Can of Worms', *The Verge*, 26 April 2022. https://www.theverge.com/2022/4/26/23040879/elon-musk-twitter-plans-free-speech-bots-anonymity-algorithm-open-source (accessed 14 May 2022); and Naomi Nix and Gerry Shih, 'Elon Musk's free-speech agenda poses safety risks on global stage', *The Washington Post*, 16 May 2022. https://www.washingtonpost.com/technology/2022/05/16/twitter-elon-musk-india/ (accessed 17 May 2022).

[191] Pranshu Verma, 'Elon Musk wants 'free speech' on Twitter. But for whom?', *The Washington Post*, 6 May 2022. https://www.washingtonpost.com/technology/2022/05/06/twitter-harassment/ (accessed 13 May, 2022).

[192] For a detailed critique of Musk's responses to issues relating to content moderation, see Mike Masnick, 'Elon Musk Demonstrates How Little He Understands About Content Moderation', *techdirt*, 15 April 2022. https://www.techdirt.com/2022/04/15/elon-musk-demonstrates-how-little-he-understands-about-content-moderation/ (accessed 13 May 2022). For an insightful analysis of the relationship between free speech and content moderation in the US context see Kate Klonick, 'The New Governors: The People, Rules and Processes Governing Online Speech' in *Harvard Law Review*, Vol. 131 (2018), pp. 1598–1670.

that without moderation social media "becomes a cesspool that people don't want to use", thus undercutting rather than promoting free speech.[193]

In addition, Musk's stance would impact his stated intent of boosting the monetisation of Twitter given that a loosening of content-moderation policies would put Twitter's advertising profits at risk from increasingly 'brand safety' conscious companies. As a result, Mike Proulx (the vice-president, research director at Forrester, a consumer research and consulting firm) notes that, while the proposed takeover is "touted as a battle over 'free speech' [it is] really a battle around content moderation: is it responsible or is it censorship? This leads to questions on whether Musk would address disinformation and hate speech on Twitter or enable it to further amplify in the name of 'free speech'."[194]

A tragic example of this tension can be seen in the racially motivated and livestreamed fatal shooting of ten people in a Buffalo, New York supermarket on 14 May 2022 that aimed to amplify its message by exploiting the viral power of extreme graphic violence. Corin Faife highlights that the shooter was seemingly radicalised online and motivated by the "great replacement theory" (a variation of white genocide theory arguing that white people are being dispossessed through immigration and interracial marriage), which he encountered on the online message board 4chan.[195] Twitter and other social-media platforms rushed to remove the shooter's manifesto and video of the shooting from their platforms, a link to a copy of the livestream remained on Facebook for ten hours after the attack and in this time was shared 46 000 times.[196] Using an extreme interpretation of free speech, as Musk seems to advocate would mean such material would remain online as a video showing graphic violence is not in itself illegal.[197]

Despite Musk's claims then, it seems unlikely that there will be radical changes to Twitter's content-moderation policies and community guidelines in the immediate

[193] Billy Perrigo, "The Idea Exposes His Naiveté.' Twitter Employees on Why Elon Musk is Wrong About Free Speech', *TIME*, 14 April 2022. https://time.com/6167099/twitter-employees-elon-musk-free-speech/ (accessed 12 May 2022). See also Filippo Menczer's arguments that weaker moderation policies would in fact hurt free speech by allowing the voices of real users to be drowned out by "malicious uses who manipulate Twitter through inauthentic accounts, bots and echo chambers". Filippo Menczer, 'Elon Musk is wrong: Research shows content rules on Twitter help preserve free speech from bots and other manipulation', *The Conversation*, 9 May 2022. https://theconversation.com/elon-musk-is-wrong-research-shows-content-rules-on-twitter-help-preserve-free-speech-from-bots-and-other-manipulation-182317 (accessed 13 May 2022).

[194] Aaron Hurst, 'What Elon Musk buying Twitter means for content moderation', *Information/Age*, 26 April 2022. https://www.information-age.com/what-elon-musk-buying-twitter-means-for-content-moderation-123499248/ (accessed 12 May 2022).

[195] Corin Faife, 'Elon Musk's silence on how he'd moderate the Buffalo shooting livestream is deafening', *The Verge*, 16 May 2022. https://www.theverge.com/2022/5/16/23076428/buffalo-shooting-video-elon-musk-twitter-content-moderation (accessed 17 May 2022).

[196] Drew Harwell and Will Oremus, Only 22 saw the Buffalo shooting live. Millions have seen it since', *The Washinton Post*, 16 May 2022. https://www.washingtonpost.com/technology/2022/05/16/buffalo-shooting-live-stream/ (accessed 17 May 2022).

[197] Corin Faife, 'Elon Musk's silence on how he'd moderate the Buffalo shooting livestream is deafening', *The Verge*, 16 May 2022. https://www.theverge.com/2022/5/16/23076428/buffalo-shooting-video-elon-musk-twitter-content-moderation (accessed 17 May 2022).

future. As Samidh Chakrabarti, Facebook's former head of civic integrity, cuttingly put it in a tweet on 14 April 2022 in response to Musk's comments regarding content moderation:

> ***Effective moderation is not inherently in conflict with free speech. It is required for people to feel free to speak. Anyone who doesn't get this has a high school stoner level grasp of societal issues and has never spent 5 min working on trust & safety.***

This potential takeover and these debates over content moderation do, however, highlight the potential for the social media landscape to rapidly change at short notice. More importantly, Shirin Ghaffary notes, Musk's efforts to influence how it functions and moderates its users raises questions about who should be able to control a company that holds so much power, as these debates signal how Twitter, despite Twitter's relatively small size, has become a key platform for politicians, business leaders, celebrities and journalists to amplify their messages and control their own narratives. In short, Twitter's social and political worth is more than its stock price.[198]

Community guidelines relating to hateful behaviour on TikTok

If Twitter can be considered explicitly political in its claims that it is the free speech wing of the free speech party and is loath to engage in any forms of moderation that could be perceived as censorship, TikTok sits at the other end of the spectrum. TikTok (which is owned and operated by the Chinese company ByteDance) was launched in 2017 and is a short-form video-sharing application. It is the most recent of the social-media platforms we have focused on for this report and is the only one not based in the US. Over the last two years, TikTok has been the most downloaded app in the world (reportedly crossing the 1 billion user mark in September 2021)[199] and has been described by Christopher Stokel-Walker as 'the new Facebook'. Stokel-Walker claims that "[j]ust as Facebook has shaped the internet, the ways we interact, and our approaches and attitudes to personal data for the past two decades, so TikTok has the potential to do the same for the next 20 years".[200] In South Africa, it is the fastest growing platform in the country, rising from 5 to 9 million between 2020 and 2021. Its growth has been driven by younger demographics but more people in the 25–44 age group have joined during South Africa's Covid-19 lockdowns.[201]

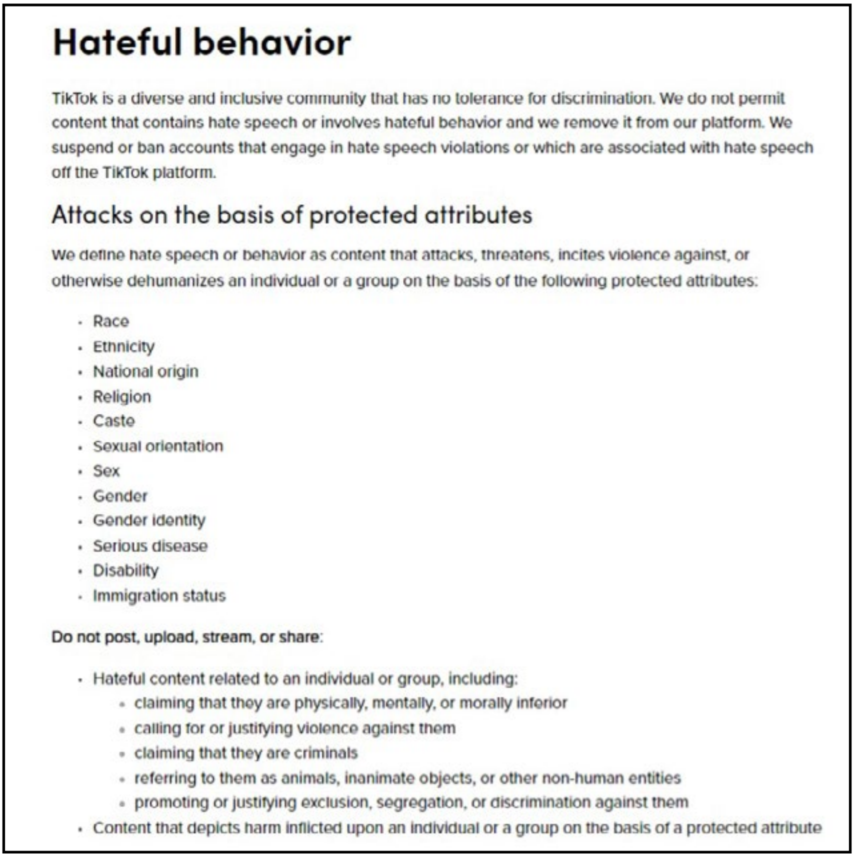
[198] Shirin Ghaffary, 'There are good reasons why Elon wants Twitter: Twitter may not be a great business, but it can buy you power and influence', Vox, 14 April 2022. https://www.vox.com/recode/23025978/elon-musk-twitter-trump-free-speech-business-facebook-youtube (accessed 26 April 2022).

[199] https://www.statista.com/statistics/1267892/tiktok-global-mau/

[200] Chris Stokel-Walker, 'TikTok is the new Facebook – and it is shaping the future of tech in its image', *The Guardian*, Monday, 16 August 2021. See also Chris Stokel-Walker, *TikTok Boom: China's Dynamite App and the Superpower Race for Social Media* (Canbury Press, 2021).

[201] Ornico and World Wide Worx, *SA Social Media Landscape Report, 2021* (2021), at page 92.

While its development has been much more rapid, much like the platforms described above, TikTok’s policies regarding what it terms ‘hateful behaviour’ has also developed as it has expanded. For example, in October 2020, it broadened its description of ‘explicitly hateful ideologies’ to include white nationalism and white genocide theory while also attempting to tackle ‘coded language and symbols that can normalise hateful speech and behaviour’.[202] These have presumably been collapsed into the category of ‘conspiring theories used to justify hateful ideologies on the community guidelines listed in Figure 3.4.2. No obviously similar category for an issue such as white genocide theory exists in Facebook and Twitter’s community guidelines, this despite the fact that Facebook claimed to have introduced a new policy banning “white nationalism“, which was almost immediately shown to have been undercut by the fact that they ignored content that did not explicitly use the term “white nationalism“ or “white separatism“.[203]



Hateful behavior

TikTok is a diverse and inclusive community that has no tolerance for discrimination. We do not permit content that contains hate speech or involves hateful behavior and we remove it from our platform. We suspend or ban accounts that engage in hate speech violations or which are associated with hate speech off the TikTok platform.

Attacks on the basis of protected attributes

We define hate speech or behavior as content that attacks, threatens, incites violence against, or otherwise dehumanizes an individual or a group on the basis of the following protected attributes:

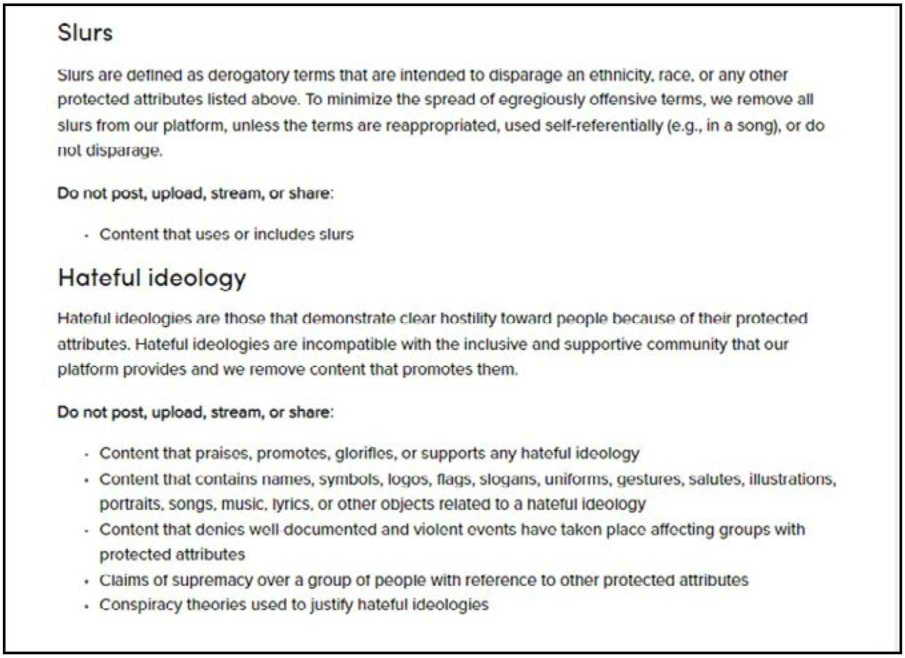
- Race
- Ethnicity
- National origin
- Religion
- Caste
- Sexual orientation
- Sex
- Gender
- Gender identity
- Serious disease
- Disability
- Immigration status

Do not post, upload, stream, or share:

- Hateful content related to an individual or group, including:
    - claiming that they are physically, mentally, or morally inferior
    - calling for or justifying violence against them
    - claiming that they are criminals
    - referring to them as animals, inanimate objects, or other non-human entities
    - promoting or justifying exclusion, segregation, or discrimination against them
- Content that depicts harm inflicted upon an individual or a group on the basis of a protected attribute

Figure 3.4.4a *TikTok Community Guidelines relating to ‘Hateful behaviour’.*

[202] Alex Hern, ‘TikTok expands hate speech ban’, *The Guardian*, 21 October 2020.
[203] Alex Hearn, Facebook ban on white nationalism too narrow, say auditors’, *The Guardian*, 1 July 2019.



Slurs

Slurs are defined as derogatory terms that are intended to disparage an ethnicity, race, or any other protected attributes listed above. To minimize the spread of egregiously offensive terms, we remove all slurs from our platform, unless the terms are reappropriated, used self-referentially (e.g., in a song), or do not disparage.

Do not post, upload, stream, or share:

- Content that uses or includes slurs

Hateful ideology

Hateful ideologies are those that demonstrate clear hostility toward people because of their protected attributes. Hateful ideologies are incompatible with the inclusive and supportive community that our platform provides and we remove content that promotes them.

Do not post, upload, stream, or share:

- Content that praises, promotes, glorifies, or supports any hateful ideology
- Content that contains names, symbols, logos, flags, slogans, uniforms, gestures, salutes, illustrations, portraits, songs, music, lyrics, or other objects related to a hateful ideology
- Content that denies well-documented and violent events have taken place affecting groups with protected attributes
- Claims of supremacy over a group of people with reference to other protected attributes
- Conspiracy theories used to justify hateful ideologies

Figure 3.4.4b *TikTok Community Guidelines relating to 'Hateful behaviour'.*

Unlike Twitter, TikTok has been upfront about its desire for the app to be a 'politics-free zone'. For example, when Raj Mishra (TikTok's head of operation in India) was asked if they would allow criticism of Indian Prime Minister Narendra Modi to be prominently featured in the app, the response was 'No' and that their ambition was to be a "one stop entertainment platform where people come to have fun rather than creating any political strife".[204] This, however, does not extend to China itself, as ByteDance CEO Zhang Yiming has stated on the record that he will ensure his products serve to promote the Chinese Communist Party's propaganda agenda.[205] In addition to this, there have been numerous claims of censorship and the curation and control of information on the site. This approach to content moderation, Stokel-Walker argues, is a legacy of the app's early development in a country with a highly controlled digital space, which stands in direct contrast to the experiences of Facebook and Twitter in the US.[206]

While their guidelines regarding 'hateful behaviour' are similar to those on Facebook and Twitter, TikTok is, Fergus Ryan et al claim, "the first globally popular social media network to take a heavy-handed approach to content moderation". They go

[204] Fergus Ryan, Audrey Fritz and Daria Impiombato, 'TikTok and WeChat curating and controlling global information flows' (Australian Strategic Policy Institute, 2020), at page 21.

[205] Ibid., at page 3.

[206] Chris Stokel-Walker, 'TikTok is the new Facebook – and it is shaping the future of tech in its image', *The Guardian*, Monday, 16 August 2021.

on to claim that by "[p]ossessing and deploying the capability to covertly control information flows, across geographical regions, topics and languages, TikTok is positioned as a powerful political actor with a global reach".[207] The targeted nature of TikTok's global censorship, used to maintain what it considers to be an apolitical stance, is thus not apolitical; but in fact makes the platform a politically powerful actor.[208]

Ryan et al's report on TikTok and WeChat suggests that numerous hashtags are suppressed on the platform, for example, hashtags related to LGBTQ+ issues are suppressed in at least eight languages, with numerous hashtags categorised as non-existent when searched for on the platform.[209] Their report also found numerous examples of content that had been suppressed and hidden from public view (shadow-banned) which, although not technically deleted, made these posts much more difficult to find, while shadow-banned users may be unaware that others cannot see their content. A report by *The Intercept* revealed that moderators were instructed to suppress posts created "by users deemed too ugly, poor, or disabled", as well as to censor political speech, which were part of wider rigid constraints that "show[s] how TikTok controls content on its platform to achieve rapid growth in the mould of a Silicon Valley start-up while simultaneously discouraging political dissent with the sort of heavy hand regularly seen in its home country of China".[210]

While Company leaders claim the platform is free of the contentious content that has come to characterise its competitors such as Facebook, Twitter and YouTube,[211] they have been increasingly critiqued for these close links to the Chinese Government. Alex Stamos, the director of the Stanford Internet Observatory and a former chief security officer at Facebook claimed that the company "is operating under a political censorship regime [...] the Chinese government has no problem telling [its companies] where they should come down in political debates".[212]

Despite concern around these links in some circles (see, for example, the US Army and Navy ban on the use of the app by its soldiers in late 2019, due to security fears),[213] the number of TikTok users has radically increased from 65 million users at the end of 2017 to an estimated one billion active users by the end of 2021. To manage these increased numbers, TikTok has had to radically scale up the number of commercial content moderators and has reportedly poached numerous content moderators from other platforms and claimed to have 10 000 content moderators

[207] Fergus Ryan, Audrey Fritz and Daria Impiombato, 'TikTok and WeChat curating and controlling global information flows' (Australian Strategic Policy Institute, 2020), at page 3.

[208] Ibid.

[209] Ibid.

[210] Sam Biddle, Paulo Victor Ribeiro & Tatiana Dias, 'Invisible Censorship: TikTok Told Moderators to Suppress Posts by "Ugly" People and the Poor to Attract New Users', *The Intercept*, 16 March 2020.

[211] Drew Harwell and Tony Room, 'Inside TikTok: A culture clash where U.S. views about censorship often were overridden by Chinese bosses, *The Washington Post*, 5 November 2019. https://www.washingtonpost.com/technology/2019/11/05/inside-tiktok-culture-clash-where-us-views-about-censorship-often-were-overridden-by-chinese-bosses/

[212] Ibid.

[213] Nicole Gaouette and Ryan Browne, 'US Army bans soldiers from using TikTok over security worries', *CNN*, 31 December 2019.

in its employ in 2020. One of these individuals who made the shift to TikTok claimed that they shifted as TikTok “hires content moderators in-house, not through a staffing agency“, “they may have better systems in place to mitigate [PTSD as a result of content seen in the process of content moderation]“, and because “there is not as much extreme content being uploaded on TikTok yet“ (which may itself be a product to their more censorious operating model),[214] with the company’s strict application of content rules “designed to protect the platform from the anger and negativity seen elsewhere on the web“.[215]

However, following this rapid increase in the number of international commercial content moderators, there have been numerous reports of complaints that content-moderation decisions are overruled by Beijing, a tension that exists as a result of the different initial ideals regarding the value of political expression and free speech online. For example, *The Guardian* reported on leaked content-moderation guidelines that barred content, relating to a specific list of twenty ‘foreign leaders or sensitive figures’, which, despite the company stating that these rules had changed, still seemed to be censored through ‘shadow-banning’ a year after the rules had purportedly changed.[216] TikTok has also produced content guidelines that are more localised for individual countries, this has actually often led to increased censorship, while the company continues to block certain hashtags from search results. For example, #acab (‘all cops are bastards’), was suppressed in the early days of the George Floyd protests, then made available after a public backlash, only to reportedly be shadow-banned again once media scrutiny had subsided. All the while, #antiacab remained readily available.[217]

There are also numerous examples of content-moderation guidelines for specific geographic areas that go far beyond local laws. For example, *The Guardian* reported how content moderation guidelines for Turkey included censoring depictions of homosexuality such as ‘holding hand’s, ‘touching’, ‘kissing’, ‘reports of homosexual groups including news, characters, music, tv shows, pictures’, as well as material ‘protecting rights of homosexuals (parade, slogan, etc.)’ and ‘promotion of homosexuality’. This seems to be part of a broader pattern of complaints about the censorship of LGBTQ+ groups.[218] Meanwhile, a *Netzpolitik* report revealed the censoring of entire hashtags, suggesting TikTok has a system of promoting and slowing down the visibility of certain content while also making some content

[214] Sam Shead, ‘TikTok is luring Facebook moderators to fill new trust and safety hubs’, 12 November 2020, *CNBC*. https://www.cnbc.com/2020/11/12/tiktok-luring-facebook-content-moderators.html

[215] Drew Harwell and Tony Room, ‘Inside TikTok: A culture clash where U.S. views about censorship often were overridden by Chinese bosses, *The Washington Post*, 5 November 2019. https://www.washingtonpost.com/technology/2019/11/05/inside-tiktok-culture-clash-where-us-views-about-censorship-often-were-overridden-by-chinese-bosses/

[216] Fergus Ryan, Audrey Fritz and Daria Impiombato, ‘TikTok and WeChat curating and controlling global information flows’ (Australian Strategic Policy Institute, 2020), at page 6. See also Alex Hern, ‘Revealed: How TikTok censors videos that do not please Beijing’, *The Guardian*, 25 September 2019.

[217] Fergus Ryan, Audrey Fritz and Daria Impiombato, ‘TikTok and WeChat curating and controlling global information flows’ (Australian Strategic Policy Institute, 2020), at page 9.

[218] Ibid., at page 10.

invisible.[219] Thus, control of what content people see is mostly in the hands of the company, which also pad feeds with content from ‘shadow accounts’ operated by company employees posing as regular users, according to leaked documents obtained by *The Intercept*, with the purpose of maintaining a steady spray of appealing content.[220]

A former content moderator for TikTok also reported that managers in the US had instructed moderators to hide videos that included any political messages or themes, not just those related to China, with these posts remaining on the users’ profile pages but prevented from being shared more widely in TikTok’s main video feed (something that TikTok also claimed had changed as it was part of an earlier, blunter, approach intended to ‘keep the app fun’).[221] The content-moderation guidelines leaked to *The Guardian* also suggest that TikTok had been censoring ‘highly controversial topics’, including topics such as ‘separatism’ and ‘conflicts between ethnic groups’ and ‘exaggerating the ethnic conflict between black and white’, among other things.[222]

The effectiveness of this attempted censorship could perhaps be questioned, as a recent report by the Institute for Strategic Dialogues has highlighted that white supremacist material is easily spread on TikTok despite it contravening its community guidelines.[223] It is unclear whether the numerous reports over the last year are a product of an attempt to minimise censorship following prior criticism, an inability to maintain effective content moderation with the rapid rise in users, poor enforcement of the community guidelines by moderators, a result of the reduced number of content moderators available during the global Covid-19 lockdowns, or (if you are particularly cynical regarding the hand of the Chinese government in ByteDance’s operations) strategically allowed to remain for various reasons.

As alluded to above, perhaps the key difference between TikTok, Facebook and Twitter is the way the Chinese Communist Party is imbricated in the running of the platform (even though the extent to which this is the case may vary across geographical areas). ByteDance’s core algorithms, as mandated by People’s Republic of China Law, has as one of its inputs Chinese Communist Party propaganda bolstering restrictions on ‘negative’ content and encouraging posts that focus on ‘Xi Jinping thought’ and ‘core socialist values’ or content that increases the ‘international influence of Chinese culture’. Ryan et al suggest that there is strong evidence that these guidelines have already informed TikTok’s global content-moderation efforts as key ByteDance executives have stated on record that they would ensure their products serve and promote the Chinese Communist Party and

[219] Chris Köver and Markus Reuter, ‘TikTok curbed reach for people with disabilities’, *Netzpolitik.org*, 2 December 2019.

[220] Sam Biddle, Paulo Victor Ribeiro & Tatiana Dias, ‘Invisible Censorship: TikTok Told Moderators to Suppress Posts by “Ugly” People and the Poor to Attract New Users’, *The Intercept*, 16 March 2020.

[221] Jack Nicas, Mike Isaac and Ana Swanson, ‘TikTok Said to be Under National Security Review’, *The New York Times*, 1 November 2019.

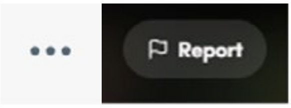
[222] Alex Hearn, ‘Revealed: How TikTok censors videos that do not please Beijing’, *The Guardian*, Wednesday, 25 September 2019.

[223] Ciaran O’Connor, *Hatescape An In-Depth Analysis of Extremism and Hate Speech on TikTok* (London: Institute for Strategic Dialogue, 2021).

that this would be integrated into the company's apps down to the algorithm level, leading to fears that feeds could be covertly tweaked by nudging content favouring certain governments.[224]

However, it is important to remember that, despite the different ideological positions of each of these platforms and the different terms used, the wording used across these platforms is often similar and contains significant overlaps. This may in part be due to an increasing convergence of views as to what is acceptable content but is also likely in part due to the exchanges between policy teams at these platforms and the movement of actual personnel across them. Some may share the same parent company (for example, Instagram and Facebook are both owned by Meta) and share resources. Many platforms also use the same companies to outsource their content moderation needs to; for example, Accenture has commercial content-moderation contracts with Facebook, YouTube, Twitter, Pinterest and others.[225]

While this section has attempted to provide a broad overview of the ideological positioning of Facebook, Twitter and TikTok and their community guidelines, as well as introducing some of the many criticisms that have been levelled at each of the platforms in order to understand some of the unique features of each platform, one of the things that unites each of them is their reliance on users as content moderators. This process requires users to flag content that they believe goes against the community guidelines before this material is then sent on to commercial content moderators for review. It is to the dynamics of this particular process across each of these three platforms that we will turn next.



3.5 Flagging Hate Speech

While it is clear algorithmic processes are playing a greater part in content moderation (particularly since the Covid-19 lockdowns of 2020), the most obvious symbol of the highly curated nature of these social-media platforms is the 'flag' or 'report' function. The term 'flag' has in fact become a ubiquitous part of our society

[224] Fergus Ryan, Audrey Fritz and Daria Impiombato, 'TikTok and WeChat Curating and controlling global information flows' (Australian Strategic Policy Institute, 2020), at pages 18–20.

[225] Adam Satariano and Mike Isaac, 'The Silent Partner Cleaning Up Facebook for $500 Million a Year', *The New York Times*, 31 August 2021.

and the ellipses next to each piece of content are now a defining symbol of the flagging process. This stage of the content-moderation process is often ignored by users and academics alike but forms a crucial part of the development of particular forms of sociality on these platforms.

The 'flag' draws users into the process of platform governance by reporting content according to the predetermined rubric of a platform's community guidelines. However, because content moderation is platform-specific, these flagging mechanisms may vary across platforms and they also often contain minor variations depending on the device on which you are using the platform. Once a post has been flagged, it is placed in a queue to be viewed by a commercial content moderator, with different queues created according to the seriousness and type of content flagged, where it is often joined by material automatically flagged by automated tools.[226] The commercial content moderator then makes a decision on whether to keep or remove the content in question.

It is important to note that certain forms of content moderation are prioritised over others. For example, material relating to child pornography will enter an expedited queue. There have also been occasions where social-media platforms have developed rapid response teams to deal with content moderation relating to a specific issue. For example, special content-moderation teams were set up to deal with various presidential elections and Facebook set up a twenty-four seven 'special operations centre' in May 2021 to respond to content posted on its platform about the Israeli-Palestinian conflict.[227] Different types of rights and different types of posts are thus policed differently (and often unevenly). The overall process of content moderation is thus privatised (and increasingly automated) and is often seen as lacking transparency.

While some platforms allow the flagging of different pieces of information (such as a post versus a profile or a page), others, such as Twitch, do not police content but user behaviour, only allowing for a user to be reported rather than a disaggregated piece of content. Just as with the flagging mechanisms, the form that moderation takes is different depending on the platform in question, but they can be summarised under these main broad categories:

- Removing content;
- Making content inaccessible or adding content warnings for content of a sensitive nature that does not quite break the rules;
- Hashtag bans that limit the search results for certain tags or show warnings when users search potentially sensitive terms. These are often not made known to those using them to avoid the coining of 'workaraound' tags. The operation of these bans is therefore opaque.

[226] Ysabel Gerrard, 'The Best-Kept Secret in Tech' in Devan Rosen (ed.), *The Social Media Debate: Unpacking the Social, Psychological, and Cultural Effects of Social Media* (New York: Routledge, 2022), at pages 84–85.

[227] Elizabeth Culliford, 'Facebook deploys special team as Israel-Gaza conflict spreads across social media', *Reuters*, 19 May 2021.

- Suspending and/or terminating an account;
- Shadow-banning. This form of content moderation has come under increasing criticism. It essentially involves preventing certain accounts and posts from showing up in recommendation systems. So, while the posted content remains on the platform, it is difficult to find. The controversy stems over the fact that users are often not told when their accounts or posts are being shadow-banned.
- Deplatforming. While each platform has its own rules and procedures, there is a great deal of overlap when it comes to the content of their community guidelines. As a result, the actions of one platform may lead to others following in their footsteps. For example, in 2018, Alex Jones (a far-right American talk show host and propagator of various racist and antisemitic conspiracy theories) was removed across multiple platforms at the same time. As Gillespie highlights, "[m]ajor platforms keep an eye on each other, and in some moments even appear to act in concert".[228]

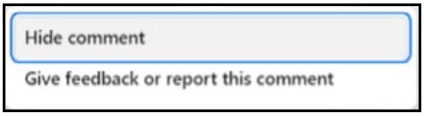
Each of these actions aligns with Western criminal justice systems that prioritise retribution and the punishment of the individual. However, Sarita Scheinebeck and Lindsay Blackwell (and the results of the research carried out by this project) argue that much of the problematic content that is posted on social media is a result of the amplification of enduring social inequities. The current social media governance approaches largely focus on simply removing individual content that violates platform policies with little to no attention given to the individuals and communities that experience the harm.[229]

Red flags and dark patterns: Flagging logics on Facebook, Twitter and TikTok

As mentioned above, each of these social-media platforms offers users the option of flagging content that they find problematic or offensive. Facebook, for example, allows you to report a profile, a post, a message, a page, or a group. With each option, a dropdown menu gives you a list of prepopulated options as to which of the community guidelines the profile has breached. Once you click the ellipsis (see Figure 3.5.1), two options appear: 'Hide comment' and 'Give feedback or report this comment'. The Hide comment option is placed first, once again suggesting Facebook's default position that you should simply ignore information you do not want to see and create a curated feed tailored to your tastes and sensitivities, while leaving other users on the platform to continue their behaviour as before.

[228] Tarleton Gillespie et al, 'Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates' in *Internet Policy Review*, Vol. 9, No. 4 (2020), pp. 1–30, at page 5. See also, Ysabel Gerrard, 'The Best-Kept Secret in Tech' in Devan Rosen (ed.), *The Social Media Debate: Unpacking the Social, Psychological, and Cultural Effects of Social Media* (New York: Routledge, 2022), at page 89.

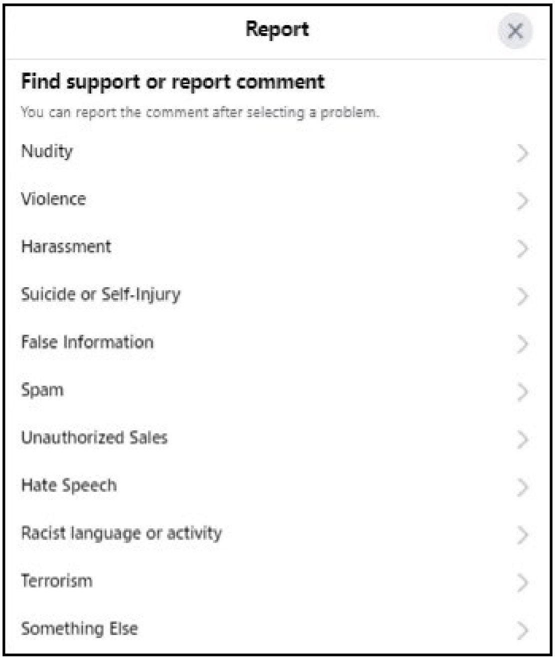
[229] See Sarita Schenebeck and Lindsay Blackwell, 'Reimagining Social Media Governance: Harm, Accountability, and Repair', *Social Science Research Network* (2021).



Hide comment

Give feedback or report this comment

Figure 3.5.1



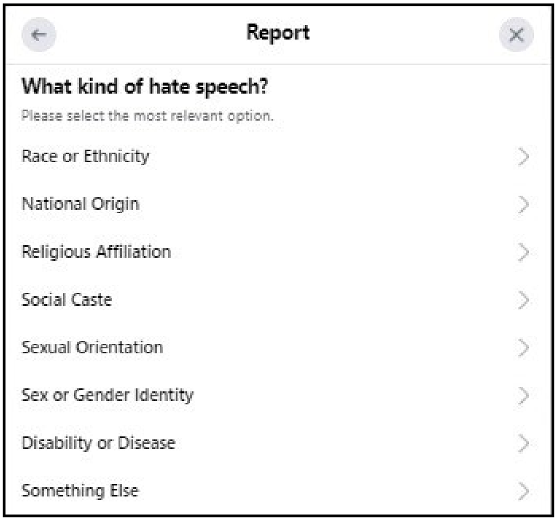
Report

Find support or report comment

You can report the comment after selecting a problem.

- Nudity
- Violence
- Harassment
- Suicide or Self-Injury
- False Information
- Spam
- Unauthorized Sales
- Hate Speech
- Racist language or activity
- Terrorism
- Something Else

Fig 3.5.2



What kind of hate speech?

Please select the most relevant option.

Race or Ethnicity

National Origin

Religious Affiliation

Social Caste

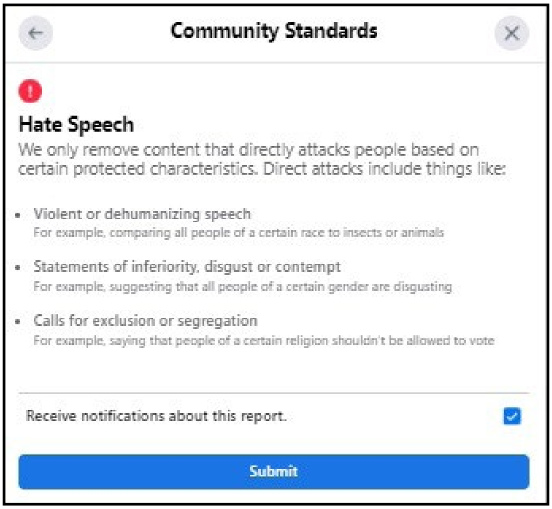
Sexual Orientation

Sex or Gender Identity

Disability or Disease

Something Else

Fig 3.5.3



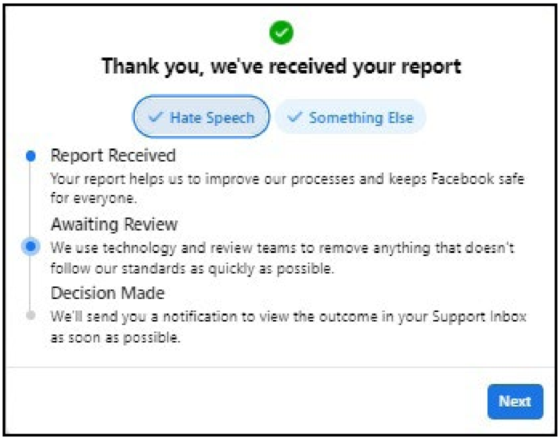
Hate Speech

We only remove content that directly attacks people based on certain protected characteristics. Direct attacks include things like:

- Violent or dehumanizing speech
  For example, comparing all people of a certain race to insects or animals
- Statements of inferiority, disgust or contempt
  For example, suggesting that all people of a certain gender are disgusting
- Calls for exclusion or segregation
  For example, saying that people of a certain religion shouldn't be allowed to vote

Receive notifications about this report. ☑

Figure 3.5.4



Thank you, we've received your report

Report Received

Your report helps us to improve our processes and keeps Facebook safe for everyone.

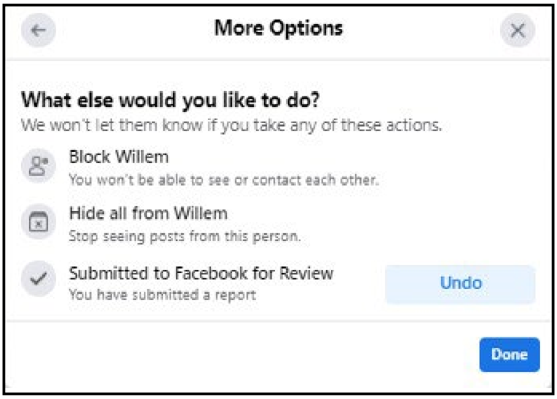
Awaiting Review

We use technology and review teams to remove anything that doesn't follow our standards as quickly as possible.

Decision Made

We'll send you a notification to view the outcome in your Support Inbox as soon as possible.

Figure 3.5.5



More Options

What else would you like to do?

We won't let them know if you take any of these actions.

Block Willem

You won't be able to see or contact each other.

Hide all from Willem

Stop seeing posts from this person.

Submitted to Facebook for Review

You have submitted a report

Figure 3.5.6

If you choose to continue and report the comment in question, a new dropdown menu appears (Figure 3.5.2), providing a series of fixed options (with more fixed options appearing if you click ‘something else’). These options are defined by the broad categories laid out in the community guidelines. When you chose ‘Hate Speech’, another dropdown menu appears (Figure 3.5.3). Once you have chosen what form of hate speech the comment in question consists of, clicking on that option sees a new page appear (Figure 3.5.4). This page is headed by a red exclamation mark and a reminder to the flagger that Facebook will only remove content that “directly attacks people based on certain protected characteristics“. The implication is clearly to get the user in question to closely consider (and perhaps reconsider) flagging the content in question.

Once you click ‘submit’ (Figure 3.5.5), a new page appears informing you that your report has been received. There is, however, one last step in the process as once you click ‘next’, you are taken to another page that not only gives you the option to undo the flag you have just created with a single click but also the options to block or hide the user whose comment you are flagging (Figure 3.5.6). Doing so means that, even if your request to remove this particular content is unsuccessful, you will no longer see material posted by that particular user. While this may be seen as a form of protecting you from further content of which you do not approve, it also means that you would be less likely to flag content from the same user again (as their content will be invisible to you).

Although there are minor differences in the format for flagging content on Twitter, it follows a similar logic where your initial attempts to flag a tweet will first give you the option to unfollow, mute, or block the account in question. If you wish to continue the reporting process, a series of options appear in a dropdown menu and two more pages of new dropdown menus with fixed options appear as you refine your reasons for reporting the content in question. On the final page of the flagging process, the following message appears: “We understand that you may not want to see every tweet and we’re sorry you saw something that offended you. Here are a few ways you can make your Twitter experience better“. This is followed by two single-click options, the one to block and the other to mute the account that produced the tweet in question.

What we see here is not only the substantially more difficult nature of flagging content than blocking or muting an account, but also the assumption that it is in fact your responsibility to ‘make your Twitter experience better’, rather than Twitter’s responsibility to ensure that you are not exposed to hateful and abusive content while they monetise your time spent on the platform and the data you produce. It is also interesting to note the implication that the tweet in question has simply offended you as an individual rather than you as a member of a particular group or category. Here we once again see that the underlying logic of these two platforms, despite their allegedly different ideological underpinnings, is to keep users (both the flagger and the individual that may be flagged) on the platform. This logic is heightened by the numerous options to censure an individual who has posted problematic content with punishments that often allow them to remain on the platform in question.

Flagging content on TikTok via a web browser on a Windows device follows a similar but more streamlined process to that described above for Facebook and Twitter. Upon hovering over the ellipsis symbol, a button with a flag and the word ‘Report’ appear. Clicking on this takes you to a dropdown menu with fixed categories developed from the various main sections of prohibited content in the community guidelines. When clicking on the ‘Hate speech’ option, a new screen appears, giving a brief summary of the kinds of content that are prohibited under the label of hate speech similar to that which we saw above. If you click ‘submit’, the process is complete.

While this streamlined process facilitates the flagging of content, it is worth noting that the ellipsis symbol for reporting content can easily be confused with the button that allows you to see comments made by other users in response to the video and add comments of your own, which consists of an ellipsis placed within a speech bubble. The flagging ellipsis is also positioned below a button that quickly and seamlessly allows you to send the content in question to friends, share it to Twitter, share it to Facebook, and share it to WhatsApp. The platform architecture, it could be argued, aims to facilitate the sharing of content ahead of the flagging of content.

This conclusion becomes more difficult to ignore when one considers that this particular emphasis of the platform architecture is even more prominent in the TikTok app for Android smartphones. Here, the ellipsis has disappeared completely, and the flagging option is instead found under the 'Share' menu (Figure 3.5.7). Once you tap on the 'Share' menu, a series of options appear (see Figure 3.5.8). These consist of one-click options to share the video in question across various social-media platforms, all highlighted in vivid colours while the dull grey and black 'Report' button can be found below these options, alongside a series of other options that allow for greater engagement with the video in question.

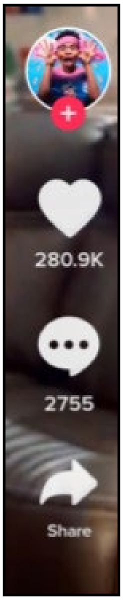


Figure 3.5.7: *Screenshot of the engagement panel of the TikTok app for Android smartphones.*

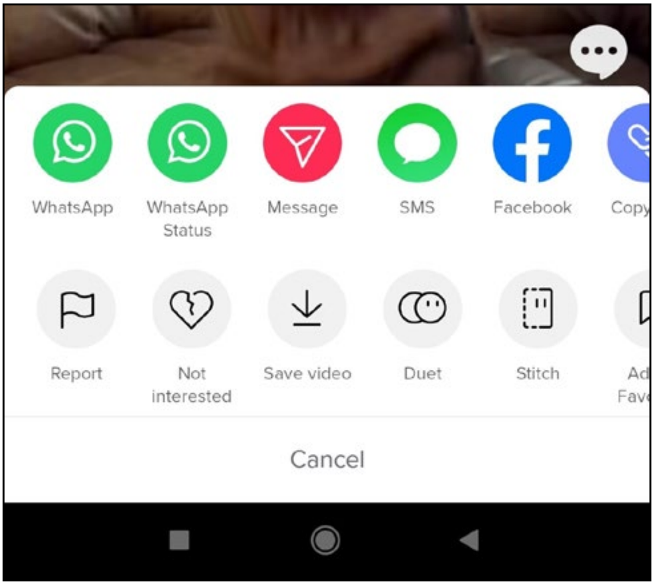


Figure 3.5.8: *Screenshot of options that appear when tapping on the 'share' button in the TikTok app on an android smartphone.*

What we can begin to see here is the way in which the intersection of computer-human interaction and platform design has an impact on flagging. The first obvious question is that of access, to put it bluntly, is the flagging button easy to find. While it could be argued that this requires some familiarity with each particular platform, social-media platforms have made it easier over time for particular posts, pages and profiles to be flagged. For example, on TikTok, the ellipses takes you to the comments made in relation to a posted video and they instead have a 'report' button listed at the bottom of the page, after all the other nudges for engagement with the content. On an Android smartphone, however, no report button is visible. Rather, you can only access the report button by pressing on the 'Share' button. Here we can already see that on the same platform, the mechanisms for reporting have slight differences depending on whether it is accessed via a Google Chrome web browser on a desktop, via a browser on a smartphone, or via the app on a smartphone.

This flagging interface on TikTok described above is a nudge towards sharing content rather than actually reporting it. A further nudge can be seen once you click on the 'share' button, which gives a range of other platforms you can share the video to, all in the vivid colours of each platform's logo while the report button is listed below them. This is part of TikTok's approach to content dissemination (which is different to other platforms), which aims to facilitate sharing in order to increase traffic to the app. As Catalina Goanta and Pietro Ortolani argue:

> ***Reports are available under this particular button, but users are nudged more towards sharing than towards filing complaints, it can be argued that the design of this button is a dark pattern. A dark pattern is an interface design choice that may result in nudging users into behaviour patterns which are against their interests.***[230]

The impact of such nudges and the relationship between users and social-media platform interfaces is beginning to attract more behavioural research, and it is clear more research is required on the topic, along with more granular data from platforms showing whether there are significant shifts in flagging practices across particular interfaces.

3.6 Flagging work

The enormous amount of user-generated content posted on social-media platforms and the significant numbers of posts that contravene community guidelines has seen these platforms turn to users themselves as a resource for flagging and facilitating the removal of such content. We have seen (and critiqued) the protocols in place for the flagging of content. This particular reliance on users to act as content moderators via these flagging protocols is a peculiarly under-researched topic.

Tarleton Gillespie and Kate Crawford are two of the few researchers who have undertaken such work and Gillespie points out how the option of flagging content, an idea that is now widespread across social media and has settled in as a norm in the logic of the social-media interface, is in fact a relatively recent introduction, with no such function being present in the earliest days of these platforms.[231] Flagging also forms part of a much broader process of user interaction on these platforms with other features such as 'likes', 'share', etc, and provides users with a sense that they are playing a part in how content on the platform is organised, ranked, valued, and presented to others. This mechanism, Crawford and Gillespie continue, thus does a considerable amount of important work "including policing content, placating users, and suggesting to external bodies that they represent a functioning system of self-regulation".[232] However, the transparency of this process varies across platforms.

[230] Catalina Goanta & Pietro Ortolani, 'Unpacking Content Moderation: The Rise of Social Media Platforms as Online Civil Courts'. Available at SSRN 3969360 (2021), at page 9.

[231] Tarleton Gillespie, 'Regulation of and by Platforms' in Jean Burgess, Alice Marwick and Thomas Poell (eds.), *The SAGE Handbook of Social Media* (London: SAGE Publications Ltd., 2018), at page 267.

[232] Kate Crawford and Tarleton Gillespie, 'What is a flag for? Social media reporting tools and the vocabulary of complaint' in *New Media & Society*, Vol. 18, No. 3 (2016), pp. 410–428, at page 418.

Crawford and Gillespie suggest that the larger platforms, such as Facebook, have come under increasing scrutiny for their content-moderation practices (or the lack thereof). For such platforms, the post-flagging process is a tactical response to these critiques. Frequently accused of being both strict, hypocritical, and unresponsive when it comes to the issue of content moderation, Facebook's introduction of a 'support dashboard' allows users to monitor the flags they have registered, view its progress, and see brief descriptions of why action was or was not taken against the flagged content. We can thus see that even across platforms with similar flagging mechanisms, there remains a spectrum of micro-practices that allow for greater or lesser articulated feedback.[233]

Some argue that flagging is a coded form of participation that claims to allow users to participate in the governance of the platform and impose (or at least attempt to impose) their ideas of what the community norms should be.[234] Others, however, argue that the interfaces often only allow for flagging under the terms and categories created by the platform and that the reporting mechanisms are not primarily designed to give users access to platform justice, but to channel the policy areas on which platforms want to take measures and to allow for content labelling that is then used in various recognition models (which once again delineate material according to these pre-determined policy areas).[235]

The limited and fixed options provided to flag content that we saw previously indicate how users are only allowed to make decisions according to the predetermined rubric of the platform's community guidelines, which as we have seen earlier are themselves often flawed and contentious. The utility of each of these predetermined options when it comes to different forms of content may lead to uneven flagging both within and across platforms. Meanwhile, the different rubrics offered, different flagging mechanisms, and different systems of human and algorithmic governance of content inevitably lead to differences in flagging practices across platforms.[236]

Due to the restrictive nature of the prescribed flagging categories, when the 'wrong' label is used for flagging content, it will likely result in an unsuccessful report. For Goanta and Ortolani, the limited choices available for reporting content via a closed number of often ill-designed and vague categories, which if not used correctly deprives flaggers of a potential remedy, suggests instead that these are designed to crowdsource algorithmic progress, increase automatically removed content, and therefore evade liability on the part of social-media platforms. This feeling is enhanced by the fact that there is often little indication on most social-media platforms regarding the process that follows flagging. The flagging process often ends with a short message stating that the report was received, while the individual whose content is removed often obtains no specific explanation as to

[233] Ibid., at page 416.

[234] Ibid., at page 411.

[235] Catalina Goanta & Pietro Ortolani, 'Unpacking Content Moderation: The Rise of Social Media Platforms as Online Civil Courts'. Available at SSRN 3969360 (2021), at page 12.

[236] Kate Crawford and Tarleton Gillespie, 'What is a flag for? Social media reporting tools and the vocabulary of complaint' in *New Media & Society*, Vol. 18, No. 3 (2016), pp. 410–428, at page 414.

why. The content itself is removed with no indication that it ever existed.[237] In short, the flagger's input in identifying problematic content is of more importance for its use in algorithmic training than in providing effective redress.[238] This seems particularly pertinent because, as we will see below, one of the key critiques of algorithmic flagging and removal is the fact that algorithms do not understand context.

Although platforms describe flagging as an expression of the community, there are questions over whether those who flag are in fact 'representative' of a larger user base, questions over who flags and why, and questions over whether such flagging is always done in good faith. Gillespie points out that platforms "are tight-lipped about how many users flag, what percentage of those who do flag provide the most flags, how often the platform decided to remove or retain content that's been flagged, etc".[239] This has not changed in the four years since Gillespie wrote this. As far as we are aware, no statistics are available on the number of flags reported on any of the social-media platforms in question for this report.

This lack of clarity can be seen as another example of the 'logic of opacity' of such platforms, but there are also numerous advantages for social-media platforms to have such a mechanism in place to begin with. First, flagging moves the burden of finding content that violates community guidelines onto the users. And secondly, Crawford and Gillespie note that "[s]ince knowledge renders the site open to liability, there is little incentive for sites to review content before users flag it".[240] This in fact was a key underpinning for the integration of user-initiated actions being embedded into the framework of social-media platforms with the flagging mechanisms becoming increasingly more visible over time.[241]

While this may initially seem contradictory, particularly when we consider the concurrent dark patterns described previously that seem to minimise the amount of content flagged, this is in fact not the case. Rather than being contradictory, this is a crucial design mechanism built into social-media platform infrastructure. Making the flagging process seem, on the face of it, extremely visible, yet having dark patterns that discourage the flagging of content baked into the platform infrastructure not only places the burden of content moderation on users while making platforms appear responsive, they at the same time limit knowledge of offensive content – thus limiting their liability.

Even when information regarding problematic content is brought to their attention, social-media platforms retain the ability to make judgments on content removal based on their own often self-interested assessments of the case at hand. There is no obligation on the part of social-media platforms to honour the flags it does

[237] Ibid., at page 415.

[238] Catalina Goanta & Pietro Ortolani, 'Unpacking Content Moderation: The Rise of Social Media Platforms as Online Civil Courts'. Available at SSRN 3969360 (2021), at page 13.

[239] Tarleton Gillespie, 'Regulation of and by Platforms' in Jean Burgess, Alice Marwick and Thomas Poell (eds.), *The SAGE Handbook of Social Media* (London: SAGE Publications Ltd., 2018), at page 268.

[240] Kate Crawford and Tarleton Gillespie, 'What is a flag for? Social media reporting tools and the vocabulary of complaint' in *New Media & Society*, Vol. 18, No. 3 (2016), pp. 410–428, at page 419.

[241] Ibid.

receive, and it is important to note that many may not in fact break the community guidelines. Given the fact that the flagging system can also be gamed (see, for example, the numerous reports of targeted flagging attacks aimed at certain groups, individuals, or content), they can also be explained away when the site prefers to ignore them.[242] Yet, at the same time, content that is removed can be done so with the justification that someone in the community complained, thus lending the removal a veneer of legitimacy. Flagged content thus acts as a means of legitimising the content curation that they undertake, while the flag system also acts as a practical and symbolic lynchpin in their much broader aim of ensuring they maintain the system of self-regulation under which they currently operate, without government (or any other) oversight.[243]

Flags , Crawford and Gillespie note, are thus not simply direct and uncomplicated representations of community sentiment and can have various meanings and functions.[244] They are often simplified versions of complex responses, all mediated by the flag interface, that are then used as simplified data points to help train the automatic flagging of other material. This is made explicitly clear in Facebook's 'how to report this' overview where each possible reporting option, once clicked, contains the benign-looking phrase "we use your feedback to help our systems learn".[245]

Social-media platforms also do not release any information relating to the proportion of flags that were or were not acted upon. We thus have very little idea of the proportion of viewers who were perturbed by a particular piece of content. Content that remains unflagged is also not necessarily unproblematic, as users may choose not to flag a particular piece of content for a variety of reasons, ranging from approval to ambivalence to disapproval but a belief that no content should be removed. Flagging can also, as mentioned above, be used tactically through coordinated and systematic flagging to silence views. For Crawford and Gillespie then, "[n]either views nor flags can be read as a clear expression of the user community as a whole".[246] Rather, they conclude, flagging and content removal in its current form "claims to proceduralise and perform collective governance while simultaneously obscuring it".[247] It is some of the issues with this obscuring nature of the removal of content that we will briefly turn to next.

[242] Ibid.

[243] Ibid., at page 412.

[244] Ibid., at page 411.

[245] 'How to Report Things', https://www.facebook.com/help/1380418588640631/?helpref=hc_fnav (accessed 5 February 2022).

[246] Kate Crawford and Tarleton Gillespie, 'What is a flag for? Social media reporting tools and the vocabulary of complaint' in *New Media & Society*, Vol. 18, No. 3 (2016), pp. 410–428, at page 420.

[247] Ibid., at page 423.

3.7 The costs of deletion

The flagging protocols described above, despite their proliferating sub-menus, simply reduce often complex content to a set of imprecise proxies that do not account for the multifaceted reasons why someone may choose to flag content or allow any community debate around content.[248] The response to the flag by the content moderators themselves (or, increasingly, algorithmic processes) is even more flattened than the options given when flagging, with moderators making a decision based on why the content was flagged, with one of only three options open to them: approve (and remove or hide the content), deny (and keep the content), or escalate (where at some point the binary decision of approve or deny will also need to be made).[249]

Certain social-media platforms are now making some effort to show flaggers the results relating to content they have flagged. Facebook, for example, has introduced a 'support dashboard' that allows users to monitor the flags they have registered, view their progress, and see brief descriptions of why action was or was not taken against the flagged content. This feedback is often perfunctory and on most social-media platforms it does not occur at all. While in theory platform-administered content moderation is not the only avenue of redress available to users (for example, in South Africa, hate speech can be tried in equality courts or reported to the South African Human Rights Commission), in practice users are unlikely to use such mechanisms for multiple reasons, ranging from cost in terms of time and finances, various bottlenecks, distrust in these institutions (a common theme in our research was a belief that the SAHRC itself was biased and incompetent), or a belief that these institutions would be unable to resolve the issue in a timeous and meaningful fashion.

It is perhaps worth reflecting here that most cases of hate speech on social media that have reached these legal platforms have been a result of a broader furore. They occurred precisely because content was not simply flagged and deleted but shared and amplified until the government felt compelled to open a case or when significant lobby groups have leant their backing (see, for example, the court cases opened by AfriForum against members of the EFF for comments made online and at political rallies).

Goanta and Ortalani note that especially when the economic value is relatively limited, potential plaintiffs remain inactive. As a result, content-moderation procedures that in comparison are relatively simple to initiate and are virtually costless to the flagger become the only available form of justice. In practice then, social-media platforms may become the only authority before which a complaint can be laid. The response to this complaint often consists of a thin form of interaction

[248] Ibid., at page 421.
[249] Ibid., at page 413.

for making a complaint (via the pre-populated options described above) and a decision being made, in a matter of seconds, by a commercial moderator who may have no sense of the context. Finally, depending on the platform, you may receive no actual confirmation of the results of your complaint unless you search out the offending content again to see if it is still available. In a context where flagging is the sole form of obtaining some form of justice, the lack of transparency becomes particularly concerning. As Goanta and Ortolani put it, "social-media platforms are the regulators, judges and enforcers of content moderation. Users present their case through procedures designed by the platform, limited by the goals of the platform, and often inaccessible to non-users".[250]

Even when a decision is made to remove content, the blunt instrument of punishment is the rendering of content or the user invisible, removing the content for everyone – not just those who were offended – and leaving no trace to allow for future debates.[251] The complete removal of content is likely due to the belief that, if the material has offended someone, it will likely offend others and once it is gone it cannot offend again, nor will it need to be adjudicated again. At the same time, it also demonstrates a commitment to protect the public while also avoiding associating the company brand with something offensive.[252]

Because of the opaque nature of content moderation on social-media platforms, users are often not aware of why they have been banned. One of the main complaints ethnographic studies relating to this issue has shown is that users are unaware as to what action triggered a sanction from the platform, an explanation of why this content was sanctioned, or who made the decision – was it algorithmic or a human moderator – and if it was a human moderator did they understand the context in which the post was made? Users are also not aware of how the action in question was flagged. A study by Nicolas P Suzor et al revealed that users often showed confusion regarding the opacity surrounding their censure by social-media platforms and that this feeling extended across political and ideological beliefs. This confusion is not helped by the often broad and vague language used in the terms of service and community guidelines, which, when combined with the lack of information regarding platforms' moderation practices, may in fact hide actual systemic bias or at the very least create apprehensions as to its existence.[253]

This lack of information often leads users to infer reasons for their sanction, leading to folk theories around what has occurred, as users develop their own rationalisation to explain the moderation decisions they were subjected to, often blaming biased moderators or systemic bias and discrimination or organised attacks

[250] Catalina Goanta & Pietro Ortolani, 'Unpacking Content Moderation: The Rise of Social Media Platforms as Online Civil Courts'. Available at SSRN 3969360 (2021), at page 7.

[251] Kate Crawford and Tarleton Gillespie, 'What is a flag for? Social media reporting tools and the vocabulary of complaint' in *New Media & Society*, Vol. 18, No. 3 (2016), pp. 410–428, at page 418 and Tarleton Gillespie, 'Regulation of and by Platforms' in Jean Burgess, Alice Marwick and Thomas Poell (eds.), *The SAGE Handbook of Social Media* (London: SAGE Publications Ltd., 2018), at page 269.

[252] Ibid.

[253] Nicolas P Suzor, Sarah Myers West, Andrew Quodling & Jillian York, 'What Do We Mean when We Talk About Transparency toward Meaningful Transparency in Commercial Content Moderation' in *International Journal of Communication,* Vol. 13 (2019), pp. 1526–1543, at page 1536.

by those who oppose their point of view. This itself often leads to more polarised views becoming entrenched. This lack of information makes it difficult for users to learn from their experience and understand the reasons for moderation, meaning the process is seen as a purely punitive one rather than a possibility to allow for user education.[254] One of the features of much of the problematic content that we encountered in our datasets was the debates that flared up around a particular piece of content. While these sometimes rapidly led to polarised discourses, there were multiple occasions when it instead led to productive debates regarding why/why not a particular piece of content was offensive. Providing a space for people to have such discussions after content has been moderated may lead to similar outcomes.

Stacking up the costs of deletion

Being banned from social media without a sense of why this has occurred may be irksome to most, but for some it can have serious consequences. Gillespie points out that, although banning by a platform cannot strictly be considered censorship, it can still have very real consequences as it can detach users from their social circle, interrupt their personal life and they would be unable to move their entire networks and personal archive of content with them when they leave. As Gillespie argues, "[t]he longer we stay on platforms and the larger they grow, the more we are compelled to stick with them and the higher the cost to leave".[255] It is thus important to ensure not only that egregious content is removed, but that those whose content has been erroneously removed have recourse to argue their case and to understand the process of what led to their content or perhaps whole profiles being removed.

It is also important to remember, however, that those who are banned – despite the clarion call of censorship – are not necessarily voiceless, as they can continue participating on other platforms and it is also worth noting that there is no legal obligation requiring social-media platforms to allow their users to speak or to restrict their users' speech.[256] A recent study by Shiza Ali et al that aimed to understand the effect of de-platforming on social networks points out how we have very little understanding of how effective the removing of content and suspending of users actually is.[257] Their study shows that being banned on one platform simply led to those users joining alternate platforms (such as Gab or Parler), where content moderation is more lax. As Assistant Professor Jeremy Blackburn, one of the members of the study, puts it: "You can't just ban these people and say, 'Hey, it worked.' They don't disappear," Blackburn said. "They go off into other places. It does have a positive effect on the original platform, but there's also some degree

[254] Ibid., at page 1531–1534.

[255] Tarleton Gillespie, 'Regulation of and by Platforms' in Jean Burgess, Alice Marwick and Thomas Poell (eds.), *The SAGE Handbook of Social Media* (London: SAGE Publications Ltd., 2018), at page 269.

[256] Ibid.

[257] Shiza Ali, Mohammad Hammas Saeed, Esraa Aldreabi, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou and Gianluca Stringhina, 'Understanding the Effect of Deplatforming on Social Networks' in *13th ACM Web Science Conference 2021* (2021), pp. 187–195.

of amplification or worsening of this type of behavior elsewhere."[258] To put it in different terms, banning often moves the content elsewhere rather than actually changing user behaviour. Quiet deletion on one platform may therefore mean very little as a censuring mechanism and instead minimises debate around content while also allowing the silent removal of content that may be so egregious that it warrants more than a quiet deletion.[259]

More concerningly, Blackburn goes on to note that those who migrated from Twitter and Reddit to alt-right and far-right platforms such as Gab and Parler tended to become more toxic, and more active with an increase in the frequency of their posts. Although these social-media platforms have a far smaller reach than Twitter and Reddit, Parler is considered to have played a significant role as an organisational tool for the 2021 United States Capitol attack. This brings forward a particularly difficult question, Blackburn notes. If by reducing these users' reach, you increase the intensity that the people who they still have access to are exposed to, is the result of their censure on more popular social-media platforms actually causing more serious real-world harm? As Blackburn puts it, "it's like a quality versus quantity type of question. Is it worse to have more people seeing this stuff? Or is it worse to have more extreme stuff being produced for fewer people?"[260]

At this point, it is worth also noting that, while all of the above has focused on social-media platforms themselves as those setting the parameters for content-moderation decisions on their platform, this is not always strictly speaking true. We have mentioned above how new forms of government regulation (or the threat thereof) have shaped content-moderation decisions but there is one important shaper of decisions that we have not yet considered, app markets, which GK Young describes as "perhaps the most influential, and often inconspicuous, parties in the social media world".[261]

We mentioned above how many of those banned from larger social networks (particularly those holding alt-right and far-right white nationalist views) had migrated to Parler, a media platform crafted in the mould of Twitter which included very few guidelines restricting user-generated content (in the process, Young notes, attempting to claim Twitter's original position as "the free speech social network").[262] Following the Capitol riots traffic on Parler dramatically increased and content moderation was not implemented by Parler itself, but by Apple, Google Play, and Amazon, which all removed the application from their stores claiming a lack of clear behavioural guidelines. Young notes that such removal (particularly

[258] Jeremy Blackburn quoted in Chris Kocher, 'Study shows users banned from social platforms go elsewhere with increased toxicity', *BingUNews*, 20 July 2021.

[259] Kate Crawford and Tarleton Gillespie, 'What is a flag for? Social media reporting tools and the vocabulary of complaint' in *New Media & Society*, Vol. 18, No. 3 (2016), pp. 410–428, at page 423.

[260] Jeremy Blackburn quoted in Chris Kocher, 'Study shows users banned from social platforms go elsewhere with increased toxicity', *BingUNews*, 20 July 2021.

[261] GK Young, 'How much is too much: The difficulties of social media content moderation' in *Information & Communications Technology Law* (2021), pp. 1–16, at page 12.

[262] Ibid., at page 13.

from Apple's iOS store, which dominates the cellular phone market, is "virtually a death sentence for a social-media platform".[263]

Following the Capitol attack, Apple demanded that Parler remove hateful and violent content from the app within 24 hours in order to remain on the App Store. When it failed to do so, it was removed. An initial attempt to be reaccepted was rejected after Apple claimed that Parler still had "highly objectionable content". It was only allowed back on the Apple Store four months later after it agreed to "more aggressively patrol what its users posted".[264] Parler is not the only social-media platform to have been impacted by content-moderation decisions made by app markets. Tumblr, a social network known for its permissiveness when it came to the posting of nudity and pornography, was removed from the iOS App Store on 16 November 2018 following the discovery of child pornography on the service. In order to return to the app store, Tumblr announced a month later that 'adult content' would no longer be allowed, which has led to its users migrating en masse to other social networks.[265]

Threats to Apple's brand management in these cases and the subsequent bans from its iOS Store led to significant changes to the content-moderation guidelines (and presumably practices) of the social-media platforms that rely on such app stores to survive. Here, content moderation took the form of the removal not of pieces of content, but of the social-media platform itself. These apps could only return after changing their policies to suit the business needs and sensibilities of app stores and is, as Gillespie notes, content moderation by other means.[266]

Gillespie refers to this process as the 'stacked' nature of content moderation, "where one intermediary must abide by the rules of another, meaning users are regulated by both together, in ways difficult to discern". Content-moderation decisions may thus be constrained by more conservative infrastructural providers (such as app stores). Such moderation at an infrastructural level, Gillespie goes on to argue, is both harder to see and to hold accountable as rules implemented further down the stack may be even less evident to users and less available for critique.[267] The act of censoring users can thus extend more broadly to the act of censoring (and deleting) social-media platforms themselves. This leads Gillespie to conclude that "[t]he moderation field is not only wide, it's also deep. Moderation decisions get made all up and down the infrastructural stack of services, often in ways that are much more opaque than the decisions made by Facebook and the like".[268]

[263] Ibid., at page 14

[264] Jack Nicas, 'Apple says Parler can return to iPhones after the app makes some changes', *The New York Times*, 19 April 2021.

[265] Jonah Engel Bromwich and Katie Van Syckle, 'Tumblr Fans Abandon Ship as Tumblr Bans Porn', *The New York Times*, 6 December 2018.

[266] Tarleton Gillespie et al, 'Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates' in *Internet Policy Review*, Vol. 9, No. 4 (2020), pp. 1–30, at page 6.

[267] Ibid., at page 7.

[268] Ibid., at page 6.

This highlights how across the stack of services that allow users access to social media, information around content-moderation practices and decisions are opaque. For users, there is little information relating to how their content was flagged, how automated detection systems that may have flagged their content work, the data they are trained on, or how they integrate with other components of the moderation system.[269] This opacity, Suzor et al note, "leads to a belief on the part of many users that their material has been removed not because it went against community guidelines, but due to a coordinated campaign and systemic bias leading to an increased distrust of both other users and of the platforms themselves".[270]

This distrust is perhaps not surprising given that critical decisions regarding content moderation are placed in the hands of a few unknown figures, who become more difficult to discern the further up the stack one travels. In the South African context, where systemic forms of bias were explicit and far-reaching, such distrust may make beliefs in systemic bias on social media particularly persuasive. Visible traces of how and why decisions were made could help avoid the appearance that one perspective in possibly conflicting world views has won and evidence of a conflict ever existing in the first place is erased.[271]

We can thus see that the flag itself, the process that led to it, and the decision of whether to remove or hide content can have serious consequences and that these decisions can be influenced by a variety of factors (though in theory are guided by the social-media platform in question's community guidelines). Given the critical importance of this decision, and despite the opacity regarding content moderators and their place in the social media infrastructure, the next part will attempt to shed some light on the conditions of those making these decisions.

3.8 Automated content moderation

We have already unpacked some of the issues around content moderation. Here, the aim is to explain the automated features of the content-moderation process in order to see how content is curated at various stages. It is important to

[269] Nicolas P Suzor, Sarah Myers West, Andrew Quodling & Jillian York, 'What Do We Mean when We Talk About Transparency toward Meaningful Transparency in Commercial Content Moderation' in *International Journal of Communication,* Vol. 13 (2019), pp. 1526–1543, at page 1536.

[270] Ibid.

[271] Kate Crawford and Tarleton Gillespie, 'What is a flag for? Social media reporting tools and the vocabulary of complaint' in *New Media & Society*, Vol. 18, No. 3 (2016), pp. 410–428, at page 423.

remember, however, that, although the moderation process is in part an attempt to prevent online harm, it is also deeply linked to the business model of these social-media platforms. Gillespie has correctly noted that Facebook (and other commercial social-media platforms) actually consists of "two intertwined networks, content and advertising, both open to all".[272] It is important then for us to think about the close links between advertising and social-media platforms and the impacts this has on content moderation.

Advertising and content moderation

In 2020, about 97.9% of Facebook's global revenue was generated from advertising, whereas only around 2% was generated by payments and other fee revenue. Facebook advertising revenue stood at close to $86 billion, with its two most important markets being the US and Canada, where average revenue per user stood at $41.41 in the last quarter of 2019 (compared to a global average of $8.52).[273] It is unsurprising then that Facebook's moderation efforts have focused on its North American markets rather than other parts of the world. Twitter generated $3.7 billion in revenue in 2020, 86% of which came from advertising (though it must be noted that it posted a net loss of $1.1 billion in 2020 and has only ever posted profits in 2018 and 2019).[274] As of the second quarter of 2021, the company reported 206 million monetisable daily active users worldwide. TikTok has been the fastest growing of the three platforms under consideration, with a revenue of $1.9 billion in 2020. TikTok's revenue streams include advertising, in-app purchases and ecommerce offerings; however, the large majority of revenues come from advertising.[275]

In a study on how a social media firm's content-moderation strategy is influenced by its revenue model, Yi Liu et al note that platforms that make their profits from advertising cannot afford to lose eyeballs or engagement on their sites. This, along with the fact that they are under public and political pressure to remove harmful content, has seen them shift to more vigorous content moderation. This shift can also be explained by the fact that the presence of problematic content may also reduce a user's enjoyment of the site, while banning users also has an effect on user engagement, which could ultimately affect the platform's profitability. Delivering eyeballs to advertisers is the most significant determinant of revenues and a mixture of content moderation with lax community standards is the best recipe to achieve this goal.[276]

[272] Tarleton Gillespie, *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media* (New Haven & London: Yale University Press, 2018)., at page 203.

[273] Revenue data taken from Statista. https://www.statista.com/statistics/271258/facebooks-advertising-revenue-worldwide/ (Accessed 22 January, 2022).

[274] Revenue data taken from Statista. https://www.statista.com/statistics/299119/twitter-net-income-quarterly/#:~:text=In%20the%20last%20reported%20quarter,the%20first%20quarter%20of%202022. (Accessed 25 August 2022).

[275] https://www.businessofapps.com/data/tik-tok-statistics/ (accessed 22 January 2022).

[276] Yi Liu, Pinar Yildirim and Z John Zhang, 'Implications of Revenue Models and Technology for Content Moderation Strategies', *Social Science Research Network* (2021).

It was in fact, to a large extent, the pressure of advertising companies that saw the scaling up of the content-moderation practices that are in operation today. Stephanie Hill notes how there were two sets of events that drove an increased focus on content-moderation policies and this opacity on the part of social platforms. The first was public hearings with representatives of social media companies over the dissemination of misinformation by Russian operatives during the 2016 election in the United States. The second was when major brands and the advertising industry found that their advertising was appearing next to distasteful, violent, or otherwise objectionable content. This resulted in several companies boycotting advertising on social media. These were two critical onslaughts as world governments, which had largely taken a light approach to regulating social-media content, were now threatening increased regulation, while advertisers (as seen above) represent the majority of social-media platforms' earnings.[277]

As a result, social media companies, particularly Facebook, announced large increases in their human content-moderation staff. They also announced the creation of 'Transparency Centres' and the increased application of automated content moderation. Social-media companies insisted that these policies were already being rolled out and therefore the problems identified by political representatives were already under control, and therefore platform self-regulation should continue to be the norm.[278] Hill notes, however, that "these developments were slow compared to changes made to meet commercial imperatives".[279]

To understand why this was the case, it is necessary to consider the history of the relationship between advertising and media. Hill (via an analysis of Harold Innis's *Empire and Communication*, a seminal work which placed the newspaper industry at the centre of US cultural imperialism) notes that once newspapers became dependent on advertising dollars, news became important in so far as it attracted readers and as a result the commercial imperatives of print media began to favour circulation over cultural or territorial integrity.[280] As a result of the dependence on advertising, a range of strategies developed to facilitate advertising. For example, the need for advertisers to reach broader audiences pushed technical developments that allowed for the printing of illustrations and the printing and shipping of more papers. Innis argued that publishing's freedom from direct control over content, in tandem with advertising interests, made circulation its priority, thus maximising its spread over geographic space. The mandates of advertising thus played a key role in defining the shape of publishing.

Hill argues that much the same is true in the case of social media, with the exception that these platforms have engineered considerable distance in many jurisdictions from laws that ordinarily hold publishers accountable for the speech present on their platforms.[281] This has been met by another significant shift, as the separation

[277] Stephanie Hill, 'Empire and the megamachine: Comparing two controversies over social media content' in *Internet Policy Review*, Vol 8, No. 1 (2019), pp. 1–18.

[278] Ibid., at page 9.

[279] Ibid., at page 2.

[280] Ibid., at pages 3–4.

[281] Ibid., at page 4.

between editorial and business concerns which were present in print media has collapsed, allowing advertisers on social media to expect not only circulation but the ability to target specific categories of users and to have advertising content integrated into the look and function of the social-media platform architectures. In direct contrast to the opacity of the transparency reports that we analyse in greater detail below, commercial incentives have become more granular and targeted with a focus on the quality of interaction with consumers rather than only circulation. As a result, there has developed an increased need on the part of advertisers to separate their content from content that might be objectionable to their target audience (what is referred to as 'brand safety').[282]

In 2017, dozens of companies, including major global advertisers, boycotted advertising on YouTube and elsewhere as advertisements had appeared next to objectionable content in what came to be referred to as the 'brand safety crisis'. Unsurprisingly, given the reliance of social-media platforms on advertising for their revenue, platforms were quick to create and implement tools to mitigate this. Advertisers could review the placement of their advertisements and the content that accompanied them, along with the promise of an increase in automated content removal, transparency centres, and human reviewers. However, unlike in the case of the public hearings, action in this instance was taken almost immediately. This included tools that changed how content on the platform was monetised along with the changes described above.

As a result of these changes, advertisements are increasingly weighted towards uncontroversial content. This does, however, raise important questions. In a world where the monetisation of content is important for a range of advocacy groups, the shift to 'brand safety' likely means that there will be less support for content around diversity and inclusion. For example, while these new measures on YouTube steered advertisers away from violent and conspiratorial content, its effects disproportionally affected educational and LGBTQ+ content, which was more likely to be flagged as 'sensational or shocking' or 'sexually suggestive' and therefore not brand safe. These policies thus worked against sexual and gender minorities. The impact of these shifts is perhaps best captured by the term that was used by users to describe them, 'the adpocalypse'.[283]

As a result of these shifts, Hill argues that what is monetisable has become the frontline of platform content governance and advertisers play a disproportionate role in defining what this process will consist of in ways that are far more immediate and far-reaching than what government has been able to achieve. It is therefore possible that the interests of advertisers can serve to curb dangerous or extreme speech on social-media platforms more effectively than governments as they lack competing interests or the need to engage with various civil society groups.[284]

However, the conservative nature of advertisers, along with their limited investment in small countries, minority populations, and political communication, means

[282] Ibid.

[283] Ibid., at pages 9–12.

[284] Ibid., at page 13.

that these issues may not be afforded the same attention as those that impact ‘brand safety’.[285] Jennifer Cobbe comes to a similar conclusion in her analysis of the increasingly important role automated content moderation is playing on social-media platforms. She notes that the commercial priorities of these platforms typically means that they intervene to suppress various forms of undesirable or unlawful communications so as to appeal to as broad a mainstream audience as possible, while also being seen to be acting responsibly for the benefit of policymakers and advertisers as their overall priorities remain primarily corporate and commercial.[286]

These threats to the self-regulation and revenue streams of social-media platforms saw a significant increase in the number of commercial content moderators employed by these platforms. In 2009, Facebook had just 12 people moderating more than 120 million users. This has expanded to an estimated 15,000 (mostly outsourced) content moderators based at content review centres across the world. This increase, along with Facebook’s focus on automated content-removal development has seen the company commit 5% of the firm’s revenue ($3.7 billion) on content moderation – an amount larger than Twitter’s entire annual revenue.[287] However, it is important to note that, while there may be a genuine desire to prevent online harm, the process is also closely tied to maintaining as many users as possible. This is highlighted by the forms of censure on these platforms, with a focus on deleting individual pieces of content rather than users wherever possible and with multiple strikes available before user bans. The economic incentives for opaque content-moderation guidelines aligned with the pressure to ‘correctly’ ban material (both in terms of removing content that can harm the ‘brand’, and in terms of not incorrectly removing content and users that can hurt the bottom line) is significant and the perception of improved content moderation is a key feature of attempting to walk this fine line.

Given the costs of content moderation and the high-profile leaks by commercial content moderators of the guidelines they are made to work with, and the increasing number of civil suits being opened against social-media platforms for harm caused to moderators in undertaking this task, there has been an increasing focus on automated content removal. Automated content moderation also became increasingly important in the immediate onset of the Coronavirus pandemic, with numerous content-moderation sites shut down due to lockdowns, resulting in an increased reliance on automated content-moderation systems.

What is automated content moderation?

In this report, we follow Ysabel Gerrard in using the term *automated content moderation* to capture the various modes of machine-based removal of content. This broader term accounts for “the range of systems designed to remove problematic

[285] Ibid.

[286] Jennifer Cobbe, ‘Algorithmic Censorship by Social Platforms: Power and Resistance’ in *Philosophy & Technology*, Vol. 34, No 7 (2021), pp. 739–766, at page 744.

[287] Yi Liu, Pinar Yildirim and Z John Zhang, ‘Implications of Revenue Models and Technology for Content Moderation Strategies’, *Social Science Research Network* (2021), at page 1.

content from social media *without direct, consistent human intervention*".[288] It thus refers to the effort to automatically prevent problematic content from ever reaching a platform or using automated tools to remove content once it has been posted if it is deemed to break the community guidelines.

There are two main modes of automated content removal. The first is 'pattern matching', where you compare new content to a list of already known examples. These include the use of 'skin filters' (which estimate the amount of skin shown in order to detect nudity) and the comparison of content to databases of previously banned images, text, videos, etc. Gillespie notes that these forms of pattern matching can only be considered Artificial Intelligence (AI) under the broadest possible definition, and the use of the term creates an aura of authority and complexity that is in fact often not present.[289] For the most part, it is quite simply the matching of a newly uploaded piece of content against an existing database of curated examples.[290] Gillespie argues that claims by Facebook in their January-March transparency report that "65.4% of [hate speech] content actioned was found and flagged by Facebook before users reported it" may sound impressive, but the overwhelming majority of what is currently being flagged is a result of pattern matching when copies of content that have already been removed by a human moderator have been matched and removed. For Gillespie, statistics like these are deliberately misleading, "implying that machine learning techniques are accurately spotting new instances of abhorrent content, not just variants of old ones."[291]

The second main form of automated content moderation is machine-learning systems. If pattern matching (as the name suggests) is aimed at matching material to pre-existing examples, machine-learning systems are trained on large datasets and aim to classify previously unseen material. In the case of social-media platforms, these datasets consist of material produced by flaggers and content moderators. The aim is to use this material to operationalise a concept like offence, abuse, or hate speech. Gorwa et al highlight that the key difference between these two forms of automated content removal is that matching requires a manual process of collating and curating individual examples of content that new material can be matched against. Classification, on the other hand, involves "inducing generalisations about features of many examples from a given category into which unknown examples may be classified (e.g. terrorist images in general)".[292] This information can then be

[288] Ysabel Gerrard, 'The Best-Kept Secret in Tech' in Devan Rosen (ed.), *The Social Media Debate: Unpacking the Social, Psychological, and Cultural Effects of Social Media* (New York: Routledge, 2022), at page 82.

[289] Tarleton Gillespie, 'Content moderation, AI, and the Question of Scale' in *Big Data and Society* (Jul.–Dec. 2020), pp. 1–5, at page 3.

[290] Robert Gorwa, Reuben Binns and Christian Katzenbach, 'Algorithmic content moderation: Technical and political challenges in the automation of platform governance', in *Big Data & Society*, Vol 7, No. 1 (2020), pp. 1–15, at page 4.

[291] Tarleton Gillespie, 'Content moderation, AI, and the Question of Scale', in *Big Data and Society* (Jul.–Dec. 2020), pp. 1–5, at page 3.

[292] Robert Gorwa, Reuben Binns and Christian Katzenbach, 'Algorithmic content moderation: Technical and political challenges in the automation of platform governance', in *Big Data & Society*, Vol 7, No. 1 (2020), pp. 1–15, at page 5.

used to down-rank certain kinds of material to reduce dissemination or to prevent material from being posted, or removing it from a platform before it is shown to other users.[293] Which of these choices will be taken is dependent on the desired governance outcome and preferences of the stakeholders that have informed the design of the system.[294]

It is important to note, however, that even machine-learning systems trained on huge datasets often still struggle to deal with complex material, as they cannot understand context. As a result, moderation is difficult to automate.[295] Though perhaps it is not the efficacy of these systems that mattered in 2016 and 2017 when social-media platforms were coming under increasing pressure, but the way in which they could be mobilised as silver bullets already available for use as a strategy of appeasement towards important stakeholders. The overhyping and mystification of automated content removal systems also allowed the presentation of "self-serving and unrealistic narratives about their [social-media platforms] technological prowess".[296] As we will see below, just how far we still have to go for these automated systems to be the silver bullet they are constantly portrayed as by social-media platforms was laid bare following the rapid and unexpected shift to automated content moderation when human moderators were sent home at the start of the Covid-19 pandemic in early 2020.

These forms of automated content moderation are far from fallible and seem to impact some demographics more than others. This is being increasingly highlighted by a range of examples regarding algorithmic bias and biased machine-learning systems that perpetuate discrimination in various forms. Given this background, Kalev Leetaru notes that it is concerning that "we are seeing precious little discussion of the impact of this bias on algorithmic content filtering."[297] Instead, the issues faced by automated content moderation once divorced from the second layer of human content moderators seems to have simply spurred on a narrative that what is required is simply better automated content removal systems rather than, for example, a much larger workforce of human content moderators.

Gorwa et al highlight how these forms of automated content removal exacerbate, rather than relieve, several key problems with content policy by focusing on three key issues. First, automated content moderation threatens to decrease understanding of an already famously opaque and secretive set of practices, thus making it even more difficult to understand and audit. It took years of mobilisation

[293] Jennifer Cobbe, 'Algorithmic Censorship by Social Platforms: Power and Resistance' in *Philosophy & Technology*, Vol. 34, No 7 (2021), pp. 739–766, at page 741.

[294] Robert Gorwa, Reuben Binns and Christian Katzenbach, 'Algorithmic content moderation: Technical and political challenges in the automation of platform governance', in *Big Data & Society*, Vol 7, No. 1 (2020), pp. 1–15, at page 6.

[295] Jennifer Cobbe, 'Algorithmic Censorship by Social Platforms: Power and Resistance' in *Philosophy & Technology*, Vol. 34, No 7 (2021), pp. 739–766, at page 741.

[296] Robert Gorwa, Reuben Binns and Christian Katzenbach, 'Algorithmic content moderation: Technical and political challenges in the automation of platform governance', in *Big Data & Society*, Vol 7, No. 1 (2020), pp. 1–15, at page 2.

[297] Kalev Leetaru, 'Twitter Follows Facebook's Dystopian Path Towards Unaccountable Automated Content Filtering', Forbes, 23 April 2019.

to get some platforms to publish (heavily edited) content-moderation guidelines and allow for any kind of independent oversight of their practices and policies. The rapid push to algorithmic moderation threatens to reverse this progress as it becomes increasingly difficult to decipher the dynamics of take-downs and the criteria by which they were made. Users have little clarity regarding whether or to what extent an automated decision led to the censure of their posted material and the databases on which these forms of automated content are trained remain closed to all despite the importance of these decisions to the landscape of free expression (and hate speech) on social media.[298]

Second, automated content-moderation systems complicate outstanding issues of justice. As mentioned above, there have recently been numerous discussions about the potential for automated content moderation to have unfair discriminatory impacts on different groups.[299] Below we will see in greater detail how this occurs in the case of language groups and may produce forms of harm against such groups. Even the act of attempting to treat all users equally (as automated content moderation is claimed to do) can have unintended consequences. Julia Angwin and Hannes Grassegger note that Facebook's algorithm is designed to defend all races and genders equally. While on the face of it this may seem a good thing, they note that this colour-blindness actually often protects those who least need it while taking it away from those who do.

This notion of equality rather than equity has much to do with the question of scale as Facebook attempts to apply consistent standards worldwide.[300] These rules fail to account for the history of discrimination and the intersectional nature of disadvantage. This is a particularly crucial issue in the South African context where toxic masculinity is often seen to dominate the public sphere and discrimination was expressly entrenched in law until 1994, with its systemic impacts still visibly prevalent and felt by the majority of the nation's non-white population.

Lastly, Gorwa et al suggest that the visibility of content is in fact a political issue and moderation itself has become a site of political contestation in many countries. Gorwa et al note that this political attention may dissipate with the rise of automated content moderation by rendering unpleasant (or perhaps even simply critical speech) largely invisible and the systems driving this themselves become hidden, much as the practices of commercial content moderation used to be. They note how, for example:

> *Facebook [could] boast of proactively removing 99.6% of terrorist propaganda [in 2019], legitimising both their technical expertise and role as a gatekeeper protecting a 'community'. However, this elides*

[298] Robert Gorwa, Reuben Binns and Christian Katzenbach, 'Algorithmic content moderation: Technical and political challenges in the automation of platform governance', in *Big Data & Society*, Vol 7, No. 1 (2020), pp. 1–15, at page 10.

[299] Ibid., at page 11. For a more detailed examination of how racism is often baked into algorithmic processes, see Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York: New York University Press, 2018).

[300] Julia Angwin and Hannes Grassegger, 'Machine Bias: Facebook's Secret Censorship Rules Protect White Men from Hate Speech but not Black Children', *ProPublica*, 28 June 2017.

> ***the hugely political question of who exactly is considered a terrorist group (Facebook only reports takedown numbers for Al-Qaeda and ISIS related content, and not for other types of terrorist content), and therefore what kind of data is trained and labelled for the classifiers, as well as the open question of the technical issues that these systems necessarily face.***[301]

They conclude that automated content removal provides a veneer of 'scientific' impartiality to the content-moderation process, allowing social-media platforms to keep the policies and data that underpin these decisions hidden while also making these decisions non-negotiable.[302] It is no surprise then that any criticism of social-media platforms is met by the promise of 'better algorithms' and 'AI', a Sirens call to the industry but one that may be no less dangerous to certain groups of users than the Sirens themselves were to sailors who encountered them. It is important to ensure that each mode of content moderation and the various ways in which they intersect, as well as their potential impacts are rigorously scrutinised.

How social-media platforms detect harmful content

While different platforms pursue different modes of content moderation, Caitlin Ring Carlson notes that the larger social-media platforms' strategies will generally consist of a mixture of three key features. First, the author of the content can self-moderate by going through the community guidelines of the platform in question. Secondly comes a process of automatic detection, which are sometimes used before, as well as after, material has been posted. Thirdly, is the process of community flagging, where users have the opportunity to flag material that they feel breaks the community guidelines. Reported content is then sent to be reviewed manually by commercial content moderators, who are often low-status, low-wage, outsourced workers in the employ of organisations dispersed globally at various worksites. These commercial content workers then decide whether to remove the material or leave it on the platform (sometimes with warnings or with down-ranked status).[303] Twitter has long been tight-lipped about the specifics of its moderation practices.[304] But it seems safe to assume, given their shift to automated tools during the Coronavirus pandemic, that they follow similar processes to those used by Facebook, and a graphic representation of Facebook's content-moderation process can be seen in Figure 3.7.1.

TikTok has also shifted to using automated content removal processes to remove videos that violate its community guidelines. According to a press release in July 2021, this process focused on an automatic content-removal system for videos

[301] Robert Gorwa, Reuben Binns and Christian Katzenbach, 'Algorithmic content moderation: Technical and political challenges in the automation of platform governance', in *Big Data & Society*, Vol 7, No. 1 (2020), pp. 1–15, at page 12.

[302] Ibid.

[303] Caitlin Ring Carlson and Hayley Rousselle, 'Report and repeat: Investigating Facebook's hate speech removal process' in *First Mondays*, Vol. 25, No. 2 (2020).

[304] Jane C Hu, 'Twitter Has Set Itself Up for an Enormous New Content Moderation Problem', *Future Tense/Slate*, 20 November 2020.

that display markers of sexual activity or nudity, violent or graphic content, as well as illegal activities and regulated goods. When it comes to potentially harmful content, videos are flagged for review and human moderators decide on whether the content should be removed or not.[305] TikTok divides material into categories of 'deleted', 'visible to self' (meaning others cannot see it), 'not recommended', and 'not for feed'. The videos in these final two categories mean the material will not be curated in the main TikTok discovery engine, which also makes the videos in question harder to find in a search.[306]

From the above, it is clear, as Robyn Caplan points out, that these platforms:

> ***[…] as businesses, technologies, and as designed spaces, make choices about the type of content that is prioritized or made visible over their networks, through algorithms, through partnerships with other companies and organizations, or through human-led curating, or through moderation. And such decisions have clear consequences for public discourse.***[307]

We have already seen how this occurs in the case of flagging by users, and in what follows we will focus on the ways in which automated content removal (and the recommender algorithms that then curate the remaining content) shapes discourse in various and sometimes perturbing ways.

Following the posting of content, automated methods are used to find and remove violative content from users' newsfeeds. Unlike with flagging, the aim is to remove content as it is uploaded to prohibit material from being seen before others are able to view, interact with, and share it.[308] Facebook has been the most visible public face of algorithmically driven content filtering, and given the number of users and the amount of content that requires moderation, Kalev Leetaru suggests that Mark Zuckerberg is in fact "betting his company's entire future on fully mechanized content filtering."[309]

[305] Nadeem Sarwar, 'TikTok's Automatic Content Removal Changes: What Creators Need To Know', *Screen Rant*, 9 July 2021.

[306] Daniel Johanson, Many are Pushing for TikTok to Release its Moderation Guidelines', *Scapi Magazine*, 24 April 2021.

[307] Robyn Caplan, 'The Artisan and the Decision Factory: The Organizational Dynamics of Private Speech Governance' in Lucy Bernholz, Hélène Landemore & Rob Reich (eds.), *Digital Technology and Democratic Theory* (Chicago: University of Chicago Press, 2021), at page 182.

[308] GK Young, 'How much is too much: The difficulties of social media content moderation' in *Information & Communications Technology Law* (2021), pp. 1–16, at page 10.

[309] Kalev Leetaru, 'Twitter Follows Facebook's Dystopian Path Towards Unaccountable Automated Content Filtering', *Forbes*, 23 April 2019.

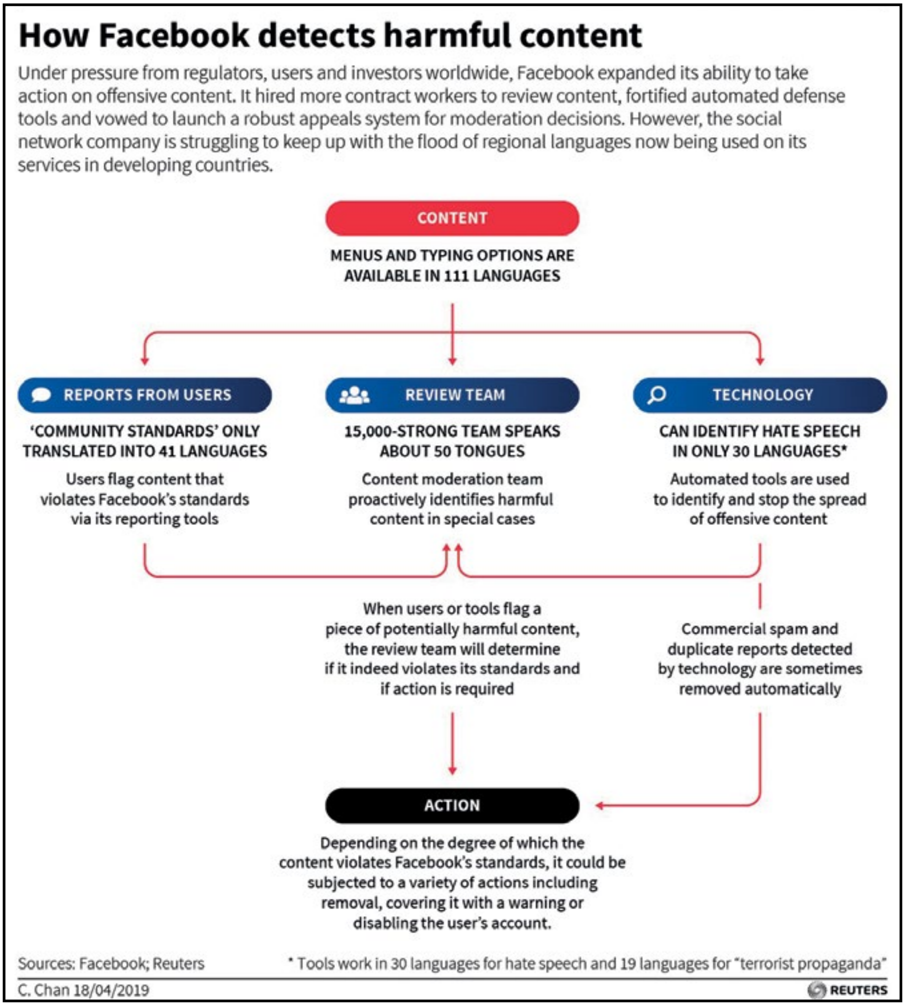

How Facebook detects harmful content
Under pressure from regulators, users and investors worldwide, Facebook expanded its ability to take action on offensive content. It hired more contract workers to review content, fortified automated defense tools and vowed to launch a robust appeals system for moderation decisions. However, the social network company is struggling to keep up with the flood of regional languages now being used on its services in developing countries.
CONTENT
MENUS AND TYPING OPTIONS ARE AVAILABLE IN 111 LANGUAGES
REPORTS FROM USERS
'COMMUNITY STANDARDS' ONLY TRANSLATED INTO 41 LANGUAGES
Users flag content that violates Facebook's standards via its reporting tools
REVIEW TEAM
15,000-STRONG TEAM SPEAKS ABOUT 50 TONGUES
Content moderation team proactively identifies harmful content in special cases
TECHNOLOGY
CAN IDENTIFY HATE SPEECH IN ONLY 30 LANGUAGES*
Automated tools are used to identify and stop the spread of offensive content
When users or tools flag a piece of potentially harmful content, the review team will determine if it indeed violates its standards and if action is required
Commercial spam and duplicate reports detected by technology are sometimes removed automatically
ACTION
Depending on the degree of which the content violates Facebook's standards, it could be subjected to a variety of actions including removal, covering it with a warning or disabling the user's account.
Sources: Facebook; Reuters
* Tools work in 30 languages for hate speech and 19 languages for "terrorist propaganda"
C. Chan 18/04/2019
REUTERS

***Figure 3.8.1:** Graphic representation of Facebook's content-moderation process.*[310]

Leetaru goes on to point out why the idea of using machines to autonomously filter content is such an alluring one: the systems never tire; can review every post with equal accuracy; would allow new content rules to be pushed out at a keystroke; remove the cost of human content moderators and with it the leaking of moderation guidelines, and they would remove the legal costs incurred in the increasingly numerous court cases being opened against social-media platforms

[310] Maggie Fick and Paresh Dave 'Facebook's flood of languages leave it struggling to monitor content', *Reuters*, 23 April 2019.

by commercial content moderators.[311] While the promises and celebration of automatic content moderation are in some ways aimed at users, lawmakers and investors, they are also, Gillespie (quoting Geiger) notes, a product of "[a] mindset prevalent in Silicon Valley, which sees these problems as technological ones requiring technological solutions".[312]

While numerous technology companies have attempted to deploy automated content-moderation tools, Facebook has been the most aggressive in its implementation and claims of its success. These claims have become an increasingly visible feature of Facebook's transparency reports. However, most researchers are sceptical that automated content moderation can succeed when it comes to contextual and cultural content, as they lack understanding of a post's intention, context, or idiom.[313] At the same time, others claim that their use will lead to the 'overcensoring' of speech.[314]

How social-media platforms ignore harmful content

Even if these forms of automated content removal were successful, there is a critical issue at play here. When it comes to written content, classification systems need data in the language in which it is being tasked to moderate. The issue of language itself is an important one in broader terms as well. A Reuters report by Maggie Fick and Paresh Dave has highlighted the extent of this issue. In it, they note that Facebook has struggled to keep up with the flood of new languages used on the platform as the rapid spread of mobile phones sees the number of people accessing social media rapidly increase all across the world.

Facebook's community guidelines – the detailed rules that act as the first step of the content-moderation process by informing users what they can and cannot post in the hope that they self-moderate – were translated into only 41 of the 111 languages supported on Facebook in March 2019. This means an estimated 652 million people worldwide speak languages supported by Facebook where the community guidelines are not translated, while an additional 230 million users or more speak one of the 31 languages that are used on Facebook but do not have official support.[315] Similar issues can be seen with other social-media platforms, with YouTube presenting guidelines in 40 of 80 available languages on its platform while Twitter's rules are available in 37 of 47 supported languages, and there is no clear indication of how many unsupported languages are used on Twitter (see

[311] Kalev Leetaru, 'Twitter Follows Facebook's Dystopian Path towards Unaccountable Automated Content Filtering', *Forbes*, 23 April 2019.

[312] Tarleton Gillespie, 'Content moderation, AI, and the Question of Scale' in *Big Data and Society* (Jul.–Dec. 2020), pp. 1–5, at page 2.

[313] GK Young, 'How much is too much: The difficulties of social media content moderation' in *Information & Communications Technology Law* (2021), pp. 1–16, at page 10.

[314] Robyn Caplan, 'The Artisan and the Decision Factory: The Organizational Dynamics of Private Speech Governance' in Lucy Bernholz, Hélène Landemore & Rob Reich (eds.), *Digital Technology and Democratic Theory* (Chicago: University of Chicago Press, 2021), at page 180.

[315] Maggie Fick and Paresh Dave, 'Facebook's flood of languages leave it struggling to monitor content', *Reuters*, 23 April 2019.

Figure 3.7.2).[316] So while what follows focuses on Facebook, other social-media platforms that are rapidly expanding are sure to face the same issues and, given Facebook's growth has been more gradual than social-media platforms like TikTok, these are likely to be even more pronounced than those encountered by Facebook.

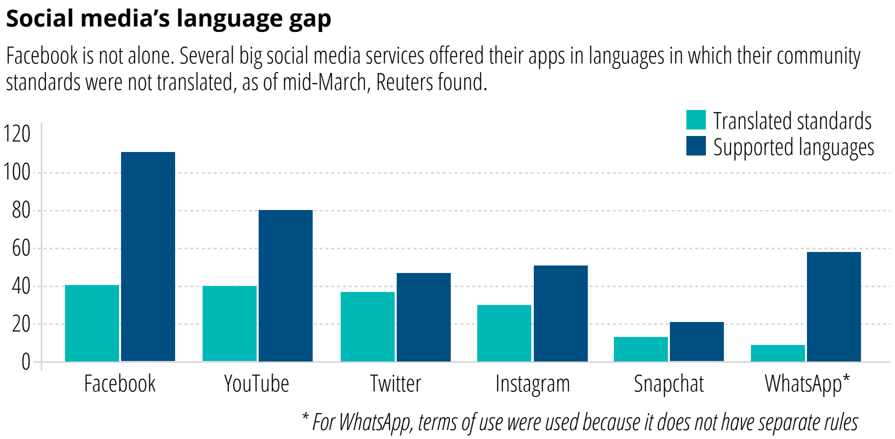


***Figure 3.8.2:** Social media's language gap.*[317]

These language gaps can have serious consequences. Facebook's determination to constantly expand before developing the expertise to properly introduce the technical and human systems that they offer to the Global North has led to accusations that Facebook helped foster ethnic cleansing in Myanmar as it was slow to add moderation tools for, and staff who could speak, Burmese.[318] This seems to be a lesson not learned, as the same processes seem at risk of repeating themselves in strife-torn regions where Facebook often dominates social media. As seen above, in some such regions, Facebook has provided menu and typing options in local languages without even having translated the community guidelines into these languages, never mind providing rigorous content moderation. To emphasise this point, Frances Haugen, a former Meta/Facebook employee who blew the whistle on the company's content-moderation practices, claimed that the company is fanning ethnic violence in Ethiopia. She went on to claim in testimony before US lawmakers that, although only 9% of Facebook users spoke English, 87% of the platform's misinformation spending was devoted to English speakers.[319] In addition, Haugen claims that Facebook has pushed into other parts of the developing world without investing in comparable protection.[320]

[316] Ibid.

[317] Ibid.

[318] Ibid.

[319] Dan Milmo, Facebook owner to 'assess feasibility' of hate speech study in Ethiopia, *The Guardian*, 14 January 2022.

[320] Cat Zakrzewski, Gerrit De Vynck, Niha Masih and Shibani Mahtani, 'How Facebook neglected the rest of the world, fueling hate speech and violence in India', *The Washington Post*, 24 October 2021.

This perhaps puts into more cautious context Facebook's claims relating to 'proactive moderation', the term Facebook uses in its transparency reports to describe content that has been removed or censured before other users have seen it. The claims made in these reports are opaque, as they report the average 'proactive removal' rates, which may mask lower rates in some languages compared to others and Facebook has declined to disclose the success rate of individual language algorithms.[321] This data of course also does not reflect on the zero percent success rate of automated content moderation in languages where no automated content classifiers have been developed. Facebook's latest transparency report makes the claim that only 3.5% of content actioned between July and September 2021 was not 'proactively' found and removed.

These figures should also give us cause for concern. If the majority of Facebook's problematic content is 'proactively' removed, and yet the majority of the languages spoken by its users are not actually supported by automated content moderation, the incredibly low figure for flagging content suggests that only a very small percentage of material has been flagged at all (which perhaps should not surprise us if there is no sense of what constitutes flaggable content given the lack of community guidelines in these languages). This suggests a failure of the content-moderation system from the beginning (the lack of content-moderation guidelines) to the end (the flagging of problematic content by other language speakers). Even in cases where problematic content has been flagged by users, little is known about discrepancies between languages when it comes to how quickly and efficiently hateful posts are removed.[322]

Facebook's skewed funding and focus, which is in large part shaped by its reliance on advertisers whose content targets consumers in the Global North, has seen some language groups better insulated against problematic content due to the increased number of content moderators who understand the language in question and as a result provide more data for machine-learning algorithms. In English, which has 1.5 billion speakers, producing datasets that number in the hundreds of thousands are produced for such training. For smaller language groups, there would be much less data to work with, something exacerbated by the fact that many hateful posts in these languages may not in fact have been flagged in the first place and if they had, may not have been removed due to the possible lack of language skills and contextual knowledge of the commercial content moderator to whom the flag was sent for adjudication.[323]

These issues obviously have a particular relevance in Africa, where Ebele Okobi, Facebook's director of public policy for Africa, told Reuters in March 2019 that the continent had the world's lowest rates of user reporting, with many people not knowing that there are community standards.[324] These figures are particularly

[321] Billy Perrigo, 'Facebook Says It's Removing More Hate Speech Than Ever Before. But There's a Catch', *Time Magazine*, 27 November 2019.

[322] Ibid.

[323] Ibid.

[324] Maggie Fick and Paresh Dave, 'Facebook's flood of languages leave it struggling to monitor content', *Reuters*, 23 April 2019.

concerning when one considers that a significant number in the continent, due to its colonial legacy, can understand French, English, Portuguese or Arabic. This suggests that there is a significant neglect in popularising the presence of these guidelines on the part of Facebook itself. Having said that, the choice not to flag may signal a broader belief that flagging content would be of little use. It is in fact an oft repeated feature in our datasets of someone pointing to material as hate speech yet seemingly not flagging it (or perhaps flagging it and having no action taken, which itself will lead to a decrease in the probability of someone flagging similar content in the future).

Despite the issues described above, Facebook has continued its push to expand into Africa. This push has been so effective that, in many parts of Africa, Nesrine Malik writes: "Facebook *is* the internet. Businesses and consumers depend heavily on it because access to the app and site are free on many African telecoms networks, meaning you don't need any phone credit to use it".[325] This has been by design, as in 2015 Facebook launched 'Free Basics', an internet service that gives users credit-free access to the platform and was designed to work on low-cost mobile phones and has been rolled out in 32 African countries.[326] Yet Facebook only opened its first content review centre in Africa in 2019 (in Nairobi) and promised to hire 100 people to cover all African markets.[327] Yohannes Eneyew Ayalew notes that they have not disclosed the number of hires made to date nor the number of moderators that focus on particular countries and no such information was seemingly made available on the platform to check this information.[328] Given this neglect, it is unsurprising that, by the end of 2020, the *Wall Street Journal* reported that "a Facebook team wrote that the risk of bad consequences in Ethiopia was dire [...] It said in some high-risk places like Ethiopia, "Our classifiers don't work, and we're largely blind to problems on our site."[329]

While Facebook presents initiatives like Free Basics as philanthropy, it is also a key commercial strategy to capture an untapped market, as users leave the platform in other more established markets. This initiative has gained increasing criticism, being banned in India in 2016 for violating rules of net neutrality, while others have begun to refer to it as a form of 'digital colonialism'. Ellery Biddle, advocacy director of Global Voices (a citizen media and activist group), has referred to it as "not introducing people to open internet where you can learn, create and build things.

[325] Nesrine Malik, 'How Facebook took over the internet in Africa – and changed everything', *The Guardian*, 22 January 2022.

[326] Ibid.

[327] https://www.facebook.com/FacebookAfrica/photos/a.468460976504996/2908779465806456/?type=3&eid=ARCqFfWL2gG7v9vPdhFQhZLzEajwbQIbAHf26jw7wUcJ69xysv2aZtmcud2XlJ-kVWmIi-kbWRNBXu3m (accessed 15 January 2022).

[328] Yohannes Eneyew Ayalew, 'Uprooting Hate Speech, The Challenging Task of Content Moderation in Ethiopia', *Center for International Media Assistance*, 27 April 2021. https://www.cima.ned.org/blog/uprooting-hate-speech-the-challenging-task-of-content-moderation-in-ethiopia/ (accessed 17 December 2022).

[329] Justin Scheck, Newley Purnell and Jeff Horwitz, 'Facebook Employees Flag Drug Cartels and Human Traffickers. The Company's Response Is Weak, Documents Show', *The Wall Street Journal*, 16 September 2021.

It's building this little web that turns the user into a mostly passive consumer of mostly western corporate content. That's digital colonialism."[330]

This view of social-media platforms as practising forms of digital colonialism is hardly helped by their seeming lack of willingness to introduce any changes to their processes to accommodate democratic African governments. For example, when the South African government asked Facebook to appear before Parliament's Portfolio Committee on Communications and Digital Technologies in May 2021 to engage with questions relating to protecting digital privacy following privacy changes on WhatsApp (which contained substantial differences in policies for users in Europe compared to those outside of Europe) and its role in misinformation, Facebook did not appear. The reason given for this was that other platforms, such as Twitter and Google SA, never responded to invitations to attend the committee meeting, leading Facebook to pull out as it "did not want to be the only tech company represented in Parliament".[331]

While there may indeed be a philanthropic underpinning to Free Basics, it is also a clear attempt to attract (or perhaps shanghai may be a more apt term) new users. This is being done despite the lack of language and content-moderation support that is offered to Facebook's core advertising markets in the Global North. More concerningly, this is being done despite clear evidence of the threat such actions may pose when it comes to the dissemination and amplification of hate speech on such social-media platforms and the very real and violent offline consequences this has had in places like Myanmar, Ethiopia, and India. The ways in which such content is rapidly disseminated and amplified form the focus of the next subsection.

3.9 How social media disseminates and amplifies harmful content

While we have seen how content moderation plays an important role in shaping the content that remains for us to see on social media, there is one particularly important form of 'soft moderation' that, although alluded to, has not been explicitly engaged with. That is the issue of recommender systems. Each social-media platform that formed part of this study relies on recommender algorithms to rank and recommend content. These algorithms take as input the content you

[330] Ellery Biddle quoted in Olivia Solon, 'It's digital colonialism': How Facebook's free internet service has failed its users', *The Guardian,* 27 July 2017.

[331] 'Facebook refuses to appear before SA Parliament on its own', *Fin24*, 25 May 2021.

engage with and rank material you are likely to also engage with at the top of your feed with the goal of maximising engagement (and therefore time and advertising revenue) on the platform.[332]

Jennifer Cobbe and Jatinder Singh refer to this as a 'surveillance business model', "whereby user behaviour is tracked and analysed to predict future behaviours and interests, personalise services, facilitate behaviourally targeted advertising, and grow user engagement, platform revenue, and market position".[333] This often occurs regardless of what the material being disseminated actually is, and is built on extensive data-gathering and analysis as platforms "obtain as much data as possible about as many people as possible doing as many things as possible from as many sources as possible" to produce datasets that are algorithmically analysed to spot patterns, interests, preferences, and predict future behaviour in order to deliver targeted advertising and personalising engagement.[334]

As Cobbe and Singh put it, "recommending is fundamentally about engagement in pursuit of profit" and "rather than showing people what they want to see, recommending shows people what the platform wants them to see".[335] These social-media platforms, more than any other form of media that has preceded them, have been more involved in mediating communication in a way that is constructing the everyday reality of billions by providing highly personalised and ever-changing content, in the process shaping users' subjective experiences of the world and giving the lie to the foundational social media governance claim that social-media platforms are not editors of content.[336] They provide and promote specifically selected information (via their often hidden algorithmic processes) to predict the relevance of certain information to certain users and then provide them that information. Cobbe and Singh note that they are not neutral conduits of content, but actively involved in shaping and promoting content via criteria that serve their own economic interests, thus taking an active role in content production (rather than mere dissemination).[337]

These recommender systems often work in two ways. Content-based filtering systems recommend content based on its similarity to previous content that has been engaged with, while collaborative filtering systems recommend content based on what similar users have consumed. Some platforms may use a hybrid of both methods, but the underlying logic of each is the same. Users are shown selected content "that the platform has determined might modify their predicted behaviour in some way – either to persuade them to click on advertising or other paid-for content, or to persuade them to stay engaged with the platform".[338]

[332] Filippo Menczer, 'Facebook whistleblower Frances Haugen testified that the company's algorithms are dangerous – here's how they can manipulate you', *The Conversation*, 20 September 2021.

[333] Jennifer Cobbe and Jatinder Singh, 'Regulating Recommending: Motivations, Considerations, and Principles' in *European Journal of Law and Technology*, Vol. 10, No. 3 (2019), pp. 1–49, at page 2.

[334] Ibid., at page 8.

[335] Ibid.

[336] Ibid., at page 15.

[337] Ibid., at page 31.

[338] Ibid., at page 8.

These systems are not neutral. Rather, they exist to achieve a particular outcome through algorithmic mediation. David Beer notes that they have outcomes in mind, and these are influenced by commercial or other interests and agendas (in this case, continued platform engagement and delivering behaviourally targeted advertising) and thus have the power to order the world in various ways. Beer refers to this more broadly as "the social power of algorithms".[339] In short, these systems have a desired outcome (even if they may not always necessarily achieve it). This creates particular difficulties for social media researchers, as each individual will have a very different feed, tailored to the thousands of data points that have been used to decide on what content to deliver and recommendations to make.

In Facebook, this is best seen in the News Feed (an algorithmically produced feed of content of various kinds determined to be most interesting or relevant to the user) and in the display of pages similar to those visited and/or liked by the user, as well as recommendations of people who may be known to the user but have not already been added as a friend. Facebook's News Feed has come under increasing attacks following the leaks by Frances Haugen, which showed that since 2018 Facebook has elevated posts that encourage interaction (as opposed to its earlier system that optimised for time spent on the platform – which itself was designed to deal with the increased popularity of clickbait). Following fears that users were spending too much time passively watching and reading material rather than interacting with it, Facebook shifted its algorithm to focus on 'meaningful social interactions'. This gives outsize weight to posts that sparked lots of comments and replies which, a *Washington Post* report claims, meant that posts that made people angry or offended gained greater traction, thus making Facebook a more polarising place. This was exacerbated by a decision by Facebook in 2017 to assign reaction emojis – including the angry emoji – five times the weight of a simple 'like'.[340]

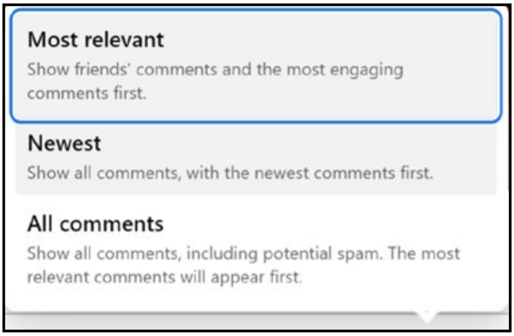
This switch prioritises posts by friends, family, and viral memes, but also divisive content. The fact that each feed reflects each user's expressed interests means that, for a subset of extremely partisan users, the algorithm can "turn their feeds into echo chambers of divisive content and news, of varying reputability, that support their outlook".[341] This is repeated across the platform architecture, with each piece of content automatically arranged to show engagement by 'most relevant' by default, which show friend's comments and the most engaging comments first (see Figure 3.8.1). Critics have argued that ordering posts from newest to oldest would not prioritise divisive content and give greater space to more frequent low-engagement posters.[342]

[339] David Beer, 'The social power of algorithms', in *Information, Communication & Society*, Vol 20, no 1, (2017), pp. 1–13. See also David Beer, *Metric Power* (London, Palgrave Macmillan, 2016).

[340] Will Oremus, Christ Alcantara, Jeremy B Merrill and Artur Galocha, 'How Facebook shapes your feed', *The Washington Post*, 26 October 2021.

[341] Ibid.

[342] Ibid.



Most relevant
Show friends' comments and the most engaging comments first.

Newest
Show all comments, with the newest comments first.

All comments
Show all comments, including potential spam. The most relevant comments will appear first.

***Figure 3.9.1:** Facebook’s arrangement of material for each piece of content.*

This deliberate “sculpting of the information landscape“, internal Facebook documents have shown, found that, for the most politically oriented 1 million American users, nearly 90% of the content shown to them was about politics and social issues. This group also received the highest amount of misinformation, with the most right-leaning of this group estimated to receive one misinformation post out of every 40.[343] Haugen has argued that we would be better off with social media feeds that simply showed us all of our friends’ posts in reverse-chronological order. This would also have uneven impacts as users and institutions that post the most frequently and with the largest existing audiences would come to dominate our feeds.[344]

On Twitter, tweets are interspersed with recommended content and recommended ‘Top Tweets’ on the user’s timeline, while trending topics are shown next to the timelines as well as suggested accounts to follow.[345] TikTok, which delivers content almost exclusively by recommender systems, has shared the broad outlines of its recommendation system, while leaked documents have provided additional insight on how it takes into account likes and comments, as well as video information like captions, sounds and hashtags. It relies heavily on how much time you spend watching each video to steer you towards other videos, often leading users down rabbit holes by driving them towards particular types of content. The document lists its ultimate goal as adding daily active users and thus optimises its metrics towards ‘retention’ (whether a user comes back for certain content) and ‘time spent’ (to keep you on the app for as long as possible). This has led to critiques that the platform’s micro-targeting of users is more likely to lead to addiction to the platform (particularly among children, a key demographic of the platform).[346]

Algorithmic personalisation via recommender systems has been the subject of increasing debate. According to leaked documents, a dummy account set up by

[343] Ibid.

[344] Ibid.

[345] Jennifer Cobbe and Jatinder Singh, ‘Regulating Recommending: Motivations, Considerations, and Principles’ in *European Journal of Law and Technology*, Vol. 10, No. 3 (2019), pp. 1–49, at page 9.

[346] Ben Smith, ‘How TikTok Reads Your Mind’, *The New York Times*, 5 December 2021.

Facebook employees in February 2019 before India's general election gives a good indication of why this is the case. The account was set up to better understand the experience of a new user in India and a profile was created of a 21-year-old woman who was a resident of North India. When violence flared in Kashmir and Indian Prime Minister Narendra Modi's campaign for re-election aimed to portray him as a nationalist strongman, the feed, without any direction from the user, began to be flooded with pro-Modi propaganda and anti-Muslim hate speech.[347] Other documents showed how political actors "spammed the social network with multiple accounts, spreading anti-Muslim messages across people's news feeds in violation of Facebook's rules".[348]

Cobbe and Singh note how examples such as this highlight the fact that the reliance on user-generated content is not the only problem with social-media platforms' business models. For them, content on its own, or content viewed by a small group of people, is unlikely to be a public policy issue. However, it becomes one when it has a large audience and is combined with other, related content to reinforce the message. They highlight that, when content is algorithmically disseminated through recommending it "(a) increases its audience, potentially significantly, and (b) typically puts it alongside other, similar content", it is at this point that such content can contribute to systemic problems. As a result, they argue, interventions that focus on the hosting of content itself often miss the issue of algorithmic dissemination.[349]

It seems inevitable, given the number of posts made on social-media platforms every day and the lack of robust forms of moderation for many of the languages social-media platforms cater for, that problematic content will escape moderation. This may be an issue when it comes to individual harassment but is not necessarily a problem at a systemic level. The rapid dissemination and amplification of that content is, particularly when we consider that the same technical systems provide the mechanism for the delivery of behaviourally targeted advertising.[350] Cobbe and Singh go on to point out that the ability of recommender systems to disseminate content, determine what and how content is recommended, and the often-dominant position of social-media platforms gives them great power and influence.[351] This is exacerbated by the fact that this power often remains hidden, with research suggesting that the majority of users may not even be aware that their Facebook News Feed is algorithmically constructed.[352]

For them, it is the dissemination and amplification of problematic content (rather than its posting) that creates systemic issues, as it takes individual content items and produces systemic societal consequences. In addition, recommender systems also

[347] Cat Zakrzewski, Gerrit De Vynck, Niha Masih and Shibani Mahtani, 'How Facebook neglected the rest of the world, fueling hate speech and violence in India', *The Washington Post*, 24 October 2021.

[348] Ibid.

[349] Jennifer Cobbe and Jatinder Singh, 'Regulating Recommending: Motivations, Considerations, and Principles' in *European Journal of Law and Technology*, Vol. 10, No. 3 (2019), pp. 1–49, at page 3.

[350] Ibid., at page 4.

[351] Ibid., at page 5.

[352] Ibid., at page 39.

place content alongside other content of a similar nature, in the process modifying users' attention and selections increasingly towards content of that nature. This further contributes to the development of problems at a systemic level. And the more users interact with this content (by viewing, liking or sharing), the more likely it is to be recommended to others.[353] They note that it is through such 'algorithmic feedback loops' that content becomes a systemic (rather than individual) problem. This feature of social media, Filippo Menczer argues, is a result of the fact that people tend to associate with similar people, leading to online neighbourhoods that are not very diverse, and this is exacerbated by the ease with which people can unfriend or unfollow those they disagree with which, as we saw earlier, is something that is in fact built into the social-media platform infrastructure. [354]

These algorithmic feedback loops are themselves amplified by the monopolistic nature of the largest social-media platforms. This prioritisation of engagement is thus likely to not only favour content that produces an emotional response, is shocking, or extreme (and numerous studies have shown how recommending can play a role in the promotion of violent extremism), but have both contributed to the growth of, and emboldened radicalised groups.[355] These recommender systems also help to shape the content produced, as users attempt to maximise their logics or to game them in various ways, which incentivises the production of certain kinds of content that are most likely to 'trend' and also creates the possibility of 'Bots' being able to artificially inflate the ranking of content to increase its dissemination and move fringe ideas into the mainstream.

An example of this can be seen in our own analysis of Operation Dudula. Accounts that were linked to this movement also played a significant role in the #PutSouthAfricansFirst movement, a hashtag that posted a wide range of xenophobic content. Here one particular account, @uLerato_pillay (created in November 2019), became the primary driver of this anti-immigrant campaign. An analysis of the hashtag by the Centre for Analytics and Behavioural Change showed that this account was backed by roughly 80 interconnected Twitter accounts that acted as an echo chamber. These accounts frequently interacted and recycled each others tweets, would push certain hashtags each day, and constantly tweeted @Julius_S_Malema to capitalise on his vast reach on social media.

From 1 March to June/July 2020, the number of mentions of the hashtag PutSouthAfricaFirst rose from roughly 100 mentions to 15,000 mentions a day. That members of this network were actively trying to get their chosen hashtags to trend can be seen by the constant use of follow and follow back campaigns and attempts to get their followers to retweet and push certain hashtags (see, for example, Figure 3.8.2 below). This network encouraged a range of similarly

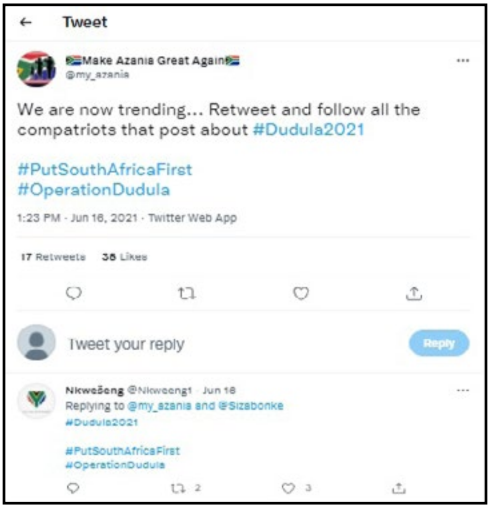
[353] Ibid., at page 17.

[354] Filippo Menczer, 'Facebook whistleblower Frances Haugen testified that the company's algorithms are dangerous – here's how they can manipulate you', *The Conversation*, 20 September 2021. https://theconversation.com/facebook-whistleblower-frances-haugen-testified-that-the-companys-algorithms-are-dangerous-heres-how-they-can-manipulate-you-169420 (accessed 13 January 2022).

[355] Jennifer Cobbe and Jatinder Singh, 'Regulating Recommending: Motivations, Considerations, and Principles' in *European Journal of Law and Technology*, Vol. 10, No. 3 (2019), pp. 1–49, at page 18.

xenophobic tweets from a range of other accounts that did not originally form part of this coordinated campaign.

When members of the same group seemingly failed to get #Operation Dudula to trend, we began to see the folk stories and conspiracy theories discussed above to explain this failure due to the opacity of the processes at play. One user (from a now-suspended account) wrote "#OperationDudula #Dudula2021 is not trending because we have been surpresed, we are being muted and silenced, MARA okusalayo Sesfikile. We Purge Illegal Foreigners, We purge ANC comrats, say no to corruption, no to incompetence, no to Unemployment, SCUM belongs _x1f6ae_ [Put Litter in Its Place Symbol]." A response to this tweet echoes the idea of suppression when the user states "I noticed as well I thought it must be a glitch or something but indeed it's totally planned,even if so it wont deter us".



🇿🇦Make Azania Great Again🇿🇦
@my_azania

We are now trending... Retweet and follow all the compatriots that post about #Dudula2021

#PutSouthAfricaFirst
#OperationDudula

1:23 PM · Jun 16, 2021 · Twitter Web App

17 Retweets 36 Likes

Nkwešeng @Nkweeng1 · Jun 16
Replying to @my_azania and @Sizabonke
#Dudula2021

#PutSouthAfricaFirst
#OperationDudula

2 3

Figure 3.9.2

These forms of recommender algorithms thus "naturally encourages and incentives the production of similar and related content, by the original producer and others".[356] Many fear that this process leads to 'filter bubbles', 'echo chambers', and polarisation.[357] It is perhaps no surprise then that the political party with some of the most radical and polarising views about race, land, and capitalism in the country (the EFF) dominated debate across our datasets, as do those accounts that push explicitly xenophobic material. Cobbe and Singh highlight, however, that recommending systems does not:

[356] Ibid., at page 19.
[357] Ibid., at page 20.

> *in and of itself cause these problems – their roots often lie in social, political, and economic causes, or just in basic human nature. But the* ***prioritisation of engagement in recommending is key to exacerbating those issues online*** *[and] can compound those underlying issues by amplifying the audience for content which is potentially problematic at a systemic level, making it easier to find other, similar content, and facilitating manipulation of the recommender systems themselves.*[358]

What this does show, however, is that the ability to promote or demote content via recommender systems positions social-media platforms as gatekeepers of political discussion (something explicitly recognised in the example above) but provides very few safeguards for when such discussions spirals into hate speech, misinformation and other forms of problematic content – particularly in the Global South. While these various forms of curating content have an impact across the world, they become particularly concerning in the African context, where social-media platforms like Facebook are often crucial sites of sometimes rapidly shrinking civil space. Yet, often the features described above exacerbate polarisation and the potential for coordinated campaigns or authoritarian governments with the will and control of limited resources to spread hateful content and disinformation to suit their political purposes.[359]

Facebook claims that the problem of content moderation is so difficult due to its gargantuan size. Its two billion users across the world make it difficult to determine what the rules of engagement on the site should be or to communicate these rules to users and moderators in a way that allows them to moderate content consistently.[360] However, size is the one thing Facebook (and most other social-media platforms) do not want to give up, something amply illustrated by Facebook's drive to increase users across the Global South despite being woefully unprepared for (or woefully apathetic to) the difficulties of content moderation in the African context.

A fundamental paradox exists here. While the content-moderation system aims to catch and remove problematic content (a process that is unevenly administered, thus causing unevenly distributed harm), their recommender systems are designed to disseminate and amplify material that is polarising (often by design) in nature and this material often consists of content that has slipped through the (sometimes non-existent) content-moderation net. The intractable problem, Jason Koebler and Joseph Cox write, is Facebook itself: "If the mission remains to connect 'everyone,' then Facebook will never solve its content moderation problem". As a result, they continue, "Facebook's content moderation team has been given a Sisyphean task: Fix the mess Facebook's worldview and business model has created, without

[358] Ibid., at pages 22–23.

[359] Tomiwa Ilori, 'Content Moderation is Particularly Hard in African Countries', *Future Tense/Slate*, 21 August 2020.

[360] Jason Koebler and Joseph Cox, 'The Impossible Job: Inside Facebook's Struggle to Moderate Two Billion People', *Vice*, 23 August 2018.

changing the worldview or business model itself".[361] It is to the harm caused to human content moderators in undertaking this Sisyphean task that we turn to next.

3.10 The implications of outsourcing content moderation

If it has become the Sisyphean task of commercial content moderators to fix the mess of Facebook's worldview due to social-media platforms' unwillingness to change their business models (which are essentially based on profit-driven and expansionist policies). These same worldviews have hobbled and harmed the content-moderation process itself. One of the key points we have attempted to make throughout this report is that the issue of content moderation is not peripheral to the task of analysing anything relating to social media, but a crucial component. The biases of content moderators and the algorithms they train no doubt shape how content moderation occurs but more importantly, the exploitative system that has been set up by these companies to deal with the central issue of content moderation (without which these platforms would unlikely be able to exist) has itself simply replicated and exacerbated the harm ordinary users experience when seeing some of the content that makes its way onto these platforms.

Unsurprisingly, the majority of research relating to the issue of commercial content moderation has focused on Facebook, by far the largest social-media platform and therefore the site with the greatest amount of content in need of moderation. However, this is also due in part to Facebook's business model, which aims to expand Facebook across the globe (even when it is not prepared for this in terms of content-moderation expertise) and to ensure that it is not barred from entering markets due to the nature of the content on the platform. Facebook thus has the highest number of content moderators. In 2019, it was reported that Facebook had a workforce of 30 000 (about 15 000 of whom were commercial content moderators), the majority of whom were hired by large professional services firms.[362]

[361] Ibid.

[362] Casey Newton, 'Bodies in Seats: At Facebook's worst-performing content moderation site in North America, one contractor has died, and others say they fear for their lives', *The Verge*, 19 June 2019. https://www.theverge.com/2019/6/19/18681845/facebook-moderator-interviews-video-trauma-ptsd-cognizant-tampa

Digital content moderation itself has now become a rapidly expanding multi-billion-dollar industry, and is projected to reach $8.8 billion in 2022 (roughly double the 2020 total) and ranges from commercial content moderation to checking for fake or duplicate user accounts, to monitoring celebrity and brand accounts to ensure they are not flooded with abuse.[363] One of the reasons little is known about this is due to the fact that the large majority of commercial content moderators are outsourced workers who are made to sign non-disclosure agreements in which they pledge not to discuss their work for Facebook. This is ostensibly meant to protect them from angry social media users and to prevent the sharing of Facebook users' personal information with the outside world.[364]

The opacity around this has meant that most material regarding the processes of commercial content moderation has had to be gleaned from the research of journalists and academics, but a spate of court cases being brought by moderators over the last two years for the harm suffered during their employment as content moderators is likely to bring much more information on these processes to light.

What is immediately clear is that the large majority of this commercial content moderation has been outsourced. The main reason for this is that outsourcing is cheaper than hiring people in-house and provides tax and regulatory benefits while also allowing for the rapid expansion and contraction of a workforce on short-term contracts, allowing for the flexibility to grow or shrink quickly depending on need. This, according to Casey Newton – a reporter who has written multiple seminal articles on the issues faced by outsourced commercial content-moderation workers in the US – has allowed Facebook to 'scale globally', and by the end of 2019 they had commercial content moderators working around the clock in more than fifty languages across more than twenty sites across the world.[365] Another reason for the outsourcing of this crucial task is the fact that it has been seen as a low-skilled job that will someday primarily be done by algorithms. This belief in a technological content-moderation utopia (despite all evidence to the contrary) therefore limits the desire to make these individuals full-time employees.[366]

This outsourcing has been done through a variety of consulting and staffing firms as well as a wider web of subcontractors, such as Accenture, that employ more than a third of Facebook's 15 000 commercial content moderators. A report by Adam Satariano and Mike Isaac has shown that in May 2021, "Accenture billed Facebook for roughly 1,900 full-time moderators in Manila; 1,300 in Mumbai, India; 850 in Lisbon; 780 in Kuala Lumpur, Malaysia; 300 in Warsaw; 300 in Mountain

[363] Adam Satariano and Mike Isaac, 'The Silent Partner Cleaning Up Facebook for $500 Million a Year', *The New York Times*, 31 August 2021. https://www.nytimes.com/2021/08/31/technology/facebook-accenture-content-moderation.html

[364] Casey Newton, 'The Trauma Floor: The secret lives of Facebook moderators in America', *The Verge*, 25 February 2019. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

[365] Ibid.

[366] Ibid.

View, Calif.; 225 in Dublin; and 135 in Austin, Texas, according to staffing records reviewed by *The [New York] Times*".[367]

The most recent addition to Facebook's outsourced commercial content-moderation stable seems to have been a centre in Nairobi opened in 2019 by Samasource (now Sama), where 200 workers from African countries have been hired. Sama, a company that refers to itself as an "ethical AI" outsourcing company headquartered in California, has contracts with Google, Microsoft and Walmart, along with Facebook (a client it does not advertise its relationship with).[368] The hub is the epicentre of the content-moderation operation for the whole of Sub-Saharan Africa and includes the moderation of content from Ethiopia, where Facebook has been used to incite violence in an increasingly brutal civil war.

This outsourcing not only saves money while tech companies report record profits, but leaves to others the work of recruiting, training and firing workers while asking these firms to achieve a range of metrics and leaving them to decide how to do this. This all occurs with the unstated threat that failure to meet perceived targets will lead to the cancellation of these multi-million-dollar contracts. According to Cori Crider, the co-founder of Foxglove (a law firm that has represented commercial content moderators in many of the court cases referred to above), this process has meant that "[e]nablers like Accenture, for eye-watering fees, have let Facebook hold the core human problem of its business at arm's length".[369] In another interview, Crider went on to state: "Outsourcing is a scam that lets Facebook rake in billions while pretending worker exploitation and union-busting is somebody else's fault [...] Foxglove has been working with Facebook moderators around the world for years – and these people have had it with exploitation, the strain of toxic content, and suppression of their right to unionize".[370]

Although Facebook proudly claims that its outsourced commercial content moderators are paid more and offered greater benefits than, for example, the larger call centre industry, the spate of court cases suggests that its workers clearly do not feel that these cover the harm suffered during their employment. While numbers regarding pay and working conditions are difficult to come across, Cognizant, which received a two-year $200 million contract from Facebook, paid its US workers $28 800 a year (roughly $2333 per month) in return for two 15-minute breaks and one 30-minute lunch break each day along with nine minutes per day of "wellness time" that they could use when they felt overwhelmed by the emotional

[367] Adam Satariano and Mike Isaac, 'The Silent Partner Cleaning Up Facebook for $500 Million a Year', *The New York Times*, 31 August 2021. https://www.nytimes.com/2021/08/31/technology/facebook-accenture-content-moderation.html

[368] Billy Perrigo, 'Inside Facebook's African Sweatshop', *Time Magazine*, 17 February 2022. https://time.com/6147458/facebook-africa-content-moderation-employee-treatment/

[369] Adam Satariano and Mike Isaac, 'The Silent Partner Cleaning Up Facebook for $500 Million a Year', *The New York Times*, 31 August 2021. https://www.nytimes.com/2021/08/31/technology/facebook-accenture-content-moderation.html

[370] Billy Perrigo, 'Inside Facebook's African Sweatshop', *Time Magazine*, 17 February 2022. https://time.com/6147458/facebook-africa-content-moderation-employee-treatment/

toll of the job.[371] Yet some of its workers claimed that its hiring policy was “find bodies wherever you can, ask as few questions as possible, and get them into a seat on the production floor where they can start working“, with one describing it as “a sweatshop in America“ with workers constantly reminded how easily they could be replaced.[372] In contrast, those in a Kenyan content-moderation site run by Sama pay its foreign employees $528 per month, with locals earning less than two-thirds of this amount, making them the lowest paid workers for the platform anywhere in the world.[373] In both cases, it was claimed that workers were far from prepared for the work that awaited them and were not provided the mental health care expertise they required in order to complete the tasks expected of them.

In a report on commercial content moderation by Jason Koebler and Joseph Cox, they note some of the difficulties of actually doing the job of commercial content moderation, with guidelines on what to keep and remove constantly shifting. One moderator described the process as “going into an office and following binary and flow-charted enforcement policies for hours on end with little idea of the nuance of the material they are analysing“.[374] This process is a necessary one as Facebook aims to get what Newton describes as “a global army of low-paid workers to consistently apply a single set of rules; near-daily changes and clarifications to those rules; a lack of cultural or political context on the part of the moderators; missing context in posts that makes their meaning ambiguous; and frequent disagreements among moderators about whether the rules should apply in individual cases“.[375]

The performance of these commercial content moderators is measured through a metric Facebook refers to as “accuracy“. This essentially means that, when Facebook audits a subset of decisions, its full-time employees must agree with the contractors. For Cognizant, one of the firms this work was outsourced to in the US, the figure required was set at 95%, with the company rarely getting close to this particular figure.[376] In other countries, it seems to be as high as 98%. As a result, the commercial content moderator must first determine whether content violates the community standards but must also select the correct reason why, as choosing the ‘wrong’ reason will count against their accuracy score. To guide them in this process, there is a 15 000-word document of internal guidelines called ‘known questions’, with new guidelines added frequently.

[371] Casey Newton, ‘Bodies in Seats: At Facebook’s worst-performing content moderation site in North America, one contractor has died, and others say they fear for their lives’, *The Verge*, 19 June 2019. https://www.theverge.com/2019/6/19/18681845/facebook-moderator-interviews-video-trauma-ptsd-cognizant-tampa

[372] Ibid.

[373] Billy Perrigo, ‘Inside Facebook’s African Sweatshop’, *Time Magazine*, 17 February 2022. https://time.com/6147458/facebook-africa-content-moderation-employee-treatment/

[374] Jason Koebler and Joseph Cox, ‘The Impossible Job: Inside Facebook’s Struggle to Moderate Two Billion People’, *Motherboard*, 23 August 2018. https://www.vice.com/en/article/xwk9zd/how-facebook-content-moderation-works

[375] Casey Newton, ‘The Trauma Floor: The secret lives of Facebook moderators in America’, *The Verge*, 25 February 2019. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

[376] Ibid.

As Timowa Ilori points out, there is also an inescapable tension at play here as platforms hope to apply global standards to content-moderation adjudication on the one hand, while deferring to local contexts on the other, claiming that they will comply with local laws upon their reviews of requests by governments.[377] However, Facebook only respects local laws when governments actively pursue their enforcement, meaning governments with the least resources are often the worst equipped to make these requests.[378] On the other side of the coin, there is the issue of changing interpretations of, for example, hate speech laws within countries (as is the case in South Africa). In addition, there is the question of what to do if the laws protect a minority with great power, as Joel Modiri suggests may well be the case when it comes to South Africa's hate speech laws.[379] In other cases, there is the question of what to do when the laws in question are incompatible with international standards.

Whatever the case, content moderators need to follow a standardised process. Once content is flagged commercial content moderators need to decide whether to leave the content online, delete it, or escalate it to another team member who may have more specialist knowledge. At times, it may even get pushed up the chain for a third review. But in addition to having to decide whether to keep or remove the material, they must make sure they choose the correct reason for removing it. When it comes to hate speech, this is a particularly difficult question. While we may all agree that some of the posts in this report constitute content that should be moderated, deciding whether this should be because it is hate speech or a slur may be a far more difficult question to answer. When it comes to hate speech, the list of possibilities for removal is longer, and it is harder to differentiate one issue from another. This means it is more difficult to pick the 'correct' answer for removal to match that of their auditor.[380] Facebook requires this data to ensure material is being removed for the "right reason" to see what is actually going on in the platform and to see what kinds of material are being successfully filtered out. This itself is a new endeavour, as *Motherboard* reports that Facebook only started collecting data on why moderators delete content in 2017.[381]

If a post contains multiple violations, moderators have to follow a hierarchy that explains which policy to delete content under. This unsurprisingly slows down the process. Those who do well as commercial content moderators are often given more hate speech cases because this is much harder to police than other types of content. However, this works against these moderators as they get scored on the

[377] Tomiwa Ilori, 'Content Moderation is Particularly Hard in African Countries', *Future Tense/Slate*, 21 August 2020. https://slate.com/technology/2020/08/social-media-content-moderation-african-nations.html

[378] Jason Koebler and Joseph Cox, 'The Impossible Job: Inside Facebook's Struggle to Moderate Two Billion People', *Motherboard*, 23 August 2018. https://www.vice.com/en/article/xwk9zd/how-facebook-content-moderation-works

[379] See Joel Modiri, 'Race, realism and critique: The politics of race and Afriforum v Malema in the (in) Equality Court: note', in *South African Law Journal*, Vol. 130 No. 2 (2013), pp. 274–293.

[380] Jason Koebler and Joseph Cox, 'The Impossible Job: Inside Facebook's Struggle to Moderate Two Billion People', *Motherboard*, 23 August 2018. https://www.vice.com/en/article/xwk9zd/how-facebook-content-moderation-works

[381] Ibid.

same basis even though they are more likely to 'make a mistake', and there are consequences if their success rate falls below a particular accuracy rate.[382]

Those interviewed by Billy Perrigo at the Sama content-moderation centre in Nairobi described how once they had logged into their station, a clock would start ticking. While sometimes the content that appeared on their screen was innocuous, at other times it might consist of graphic videos of dismemberment, murder or rape. However, no matter how disturbing the content, a decision needed to be made on it in 50 seconds (meaning 580 items per day ranging from short posts to hour-long videos needed to be moderated). This 50-second target was put in place by Sama and it could rise as high as 70 seconds or sink as low as 36 seconds depending on workload and staffing.[383] If workers take too long to make a decision on the content, they risked being reprimanded and, potentially, losing their jobs. However, if their "accuracy score" was too low, they also risked being reprimanded and potentially losing their jobs. At the same time, the outsourced company itself risks losing its contract if they do not moderate enough material with a high enough accuracy rate.

Speed, efficiency, and 'accuracy' are all built on the idea that moderators must pick "the right reason" for removing content through a notion of objectivity that is valued above all else, even though such a notion simply cannot exist when it comes to something like hate speech. Perrigo highlights that, although this pressure for speed is put in place by Sama rather than Facebook, it seems clear that the reason for doing so is a fear of losing a multimillion-dollar contract if they do not moderate enough content. This focus on speed in turn meant, as a former counsellor at Sama stated, that the moderators are not being cared for in terms of their mental health. As they put it, "Sama is more interested in productivity than the safety of the moderator".[384]

Here perhaps lies the most significant issue, that is the fact that commercial content moderators are made to constantly deal with the absolute worst that social media has to offer, ranging from dismemberment, mutilation, mass shootings, child pornography, animal abuse, hate speech and various other forms of violence when they are often not suited to the task and are not given the support they need to undertake these tasks without experiencing significant and long-lasting trauma. Such harms are repeated across the commercial content-moderation social media ecosystem as content often appears across various platforms (and is often moderated in different ways as a result). In some, this leads to Post-traumatic stress disorder (PTSD) and various other conditions that sometimes only appear long after their contract ends, while others embrace the fringe views they were hired to moderate after constantly being exposed to conspiracies at work.[385]

[382] Ibid.

[383] Billy Perrigo, 'Inside Facebook's African Sweatshop', *Time Magazine*, 17 February 2022. https://time.com/6147458/facebook-africa-content-moderation-employee-treatment/

[384] Ibid.

[385] Casey Newton, 'Bodies in Seats: At Facebook's worst-performing content moderation site in North America, one contractor has died, and others say they fear for their lives', *The Verge*, 19 June 2019.

In the US, a court case brought by commercial content moderators led to Facebook agreeing to pay $52 million to current and former moderators in the US and to provide them with greater counselling while employed.[386] Meanwhile, new cases are being opened in Europe, where more than 30 former Facebook moderators in Ireland, Spain and Germany are suing the social-media company and four of its third-party outsourcing agents after suffering psychological damage from viewing graphic content.[387] A case is also being prepared by Foxglove against Sama, a site Perrigo described as having "[a] culture characterised by mental trauma, intimidation, and alleged suppression of the right to unionize", and where numerous workers are leaving the company. Some of these workers resigned after being diagnosed with PTSD, anxiety, and depression, while others described being traumatised but unable to afford healthcare to get formal diagnoses.[388]

We described earlier how Facebook's expansionist business model has often seen the worst harms fall on the Global South, due to its inability to properly moderate the content that is posted on its platform. Now, the moderators hired to ensure that Facebook remains a viable enterprise are treated as an outsourced underclass, despite the vast profits their work helps to underwrite, with their financial and psychological wellbeing seemingly an afterthought. Once again, those bearing the worst of the brunt of this process are seemingly content moderators in the Global South. This has created a situation where not only is the deluge of hateful material posted in these nations unstemmed, but the very few hired to moderate this content are exposed to repeated harm with little to no recourse or support to deal with them. This process, Perrigo notes, "has led some observers to raise concerns that Facebook is profiting from exporting trauma along old colonial axes of power, away from the U.S. and Europe and toward the developing world".[389]

When one content moderator in the US interviewed by Newton – who described no longer sleeping for more than two or three hours a night and waking up in cold sweats, crying – was asked what he thought needed to change, the response was a terse one: "I think Facebook needs to shut down".[390] While this may not be achievable, the outsourcing of the trauma of moderation (which has been a key part of Facebook's business model) clearly needs to be rethought. Dr Paul Barret, from New York's Stern Center for Business and Human Rights, published a report in

https://www.theverge.com/2019/6/19/18681845/facebook-moderator-interviews-video-trauma-ptsd-cognizant-tampa

[386] Casey Newton, 'Facebook will pay $52 million in settlement with moderators who developed PTSD on the job', *The Verge*, 12 May 2020. https://www.theverge.com/2020/5/12/21255870/facebook-content-moderator-settlement-scola-ptsd-mental-health

[387] Nicky Harley, 'Facebook moderators across Europe sue for damages over effect of extreme content', *The National*, 10 February 2022. https://www.thenationalnews.com/world/europe/facebook-moderators-across-europe-sue-for-damages-over-effect-of-extreme-content-1.1173124

[388] Billy Perrigo, 'Inside Facebook's African Sweatshop', *Time Magazine*, 17 February 2022. https://time.com/6147458/facebook-africa-content-moderation-employee-treatment/

[389] Ibid.

[390] Casey Newton, 'Bodies in Seats: At Facebook's worst-performing content moderation site in North America, one contractor has died, and others say they fear for their lives', *The Verge*, 19 June 2019. https://www.theverge.com/2019/6/19/18681845/facebook-moderator-interviews-video-trauma-ptsd-cognizant-tampa

2020 recommending that Facebook needed to take content moderation in-house and ensure greater pay, access to medical help, and a raising of commercial content moderators' station in the workplace along with a radical increase in the number of moderators employed. As he put it, "[c]ontent moderation is not like other outsourced functions, like cooking or cleaning, it is a central function of the business of social media, and that makes it somewhat strange that it's treated as if it's peripheral or someone else's problem".[391] It seems unlikely that this will be the case in the near future, and as long as the processes described above are in place, we can expect the multiple forms of harm of social media to continue, disproportionately harming those who are least capable of withstanding their onslaughts.

3.11 The opacity of Transparency Reports

The arguments above have stressed two key points. First, that the issue of content moderation is crucial in order to understand that the datasets used for the analysis in Section Two are in fact already highly curated in various ways. Substantial amounts of content have been removed before anyone actually has a chance to see it, so what we do see is the discourse that remains after being sieved through automatic moderation processes, user sensibilities, and content-moderation decisions – all filtered via the lens of the community guidelines. Second, we have constantly returned to the issue of opacity throughout each stage of the content-moderation process and across each stack of the infrastructural network.

These two strands both come together in the 'Transparency Reports' that are increasingly being produced and regularly released by social-media platforms (as well as other internet platforms). These reports are a visible and consolidated representation of the content-moderation efforts of each of these platforms. Although they are labelled 'transparency reports', we will show that, while they are a step in the right direction, these reports encapsulate the logic of opacity of social-media platforms and the fact that they were produced only after significant pressure by various civil society groups and threats of increased government regulation suggests that this opacity is hardly accidental.

[391] Nicky Harley, 'Facebook moderators across Europe sue for damages over effect of extreme content', *The National*, 10 February 2022. https://www.thenationalnews.com/world/europe/facebook-moderators-across-europe-sue-for-damages-over-effect-of-extreme-content-1.1173124. See also Paul M Barrett, 'Who Moderates the Social Media Giants? A Call to End Outsourcing', Stern Center for Business and Human Rights. https://bhr.stern.nyu.edu/tech-content-moderation-june-2020

While some have argued that governments should take a more direct role in moderation processes on social media, the fact that these platforms are private does actually sometimes allow them to be more aggressive than a government would be when it comes to curating discourse. However, this absence of credible alternatives to content moderation has seen a push towards improved governance within these companies with a particular focus on the transparency and accountability of platforms' decision-making.[392] Social-media platforms have thus come under increasing pressure to provide more information on the scale and processes of their content-moderation practices (as well as their own commercial interests). The Global Network Initiative, a multi-stakeholder organisation that aimed to hold tech companies accountable to a set of principles, formed an important part of this process.

The Global Network Initiative's main aim is to protect and advance freedom of expression and privacy rights by providing a framework for responsible company decision-making, which included demands for public transparency regarding how much and what types of data companies turned over to various governments and how many occurrences of government censorship have been imposed on platforms. This resulted in the first publication of transparency reports by major companies such as Google in 2010. These demands expanded as civil society groups put increasing pressure on companies to provide information about how they enforce their own policies and practices.[393] This coalesced around the formation of the Santa Clara Principles on Transparency and Accountability in Content Moderation, which aimed to provide some baseline principles around content moderation that focused on three main issues:

- Companies should publish the numbers of posts removed and accounts permanently or temporarily suspended due to violations of their content guidelines.
- Companies should provide notice to each user whose content is taken down or account is suspended about the reason for the removal or suspension.
- Companies should provide a meaningful opportunity for timely appeal of any content removal or account suspension.

As a result of such pressure, social-media platforms are increasingly publishing 'transparency reports' in an attempt to lessen public concerns regarding the role they play as arbiters of speech online and increasing calls that they should be accountable for their content-moderation practices (as well as, increasingly, their need to partially fulfil legal requirements such as Germany's NetzDG legislation, which requires by law that social networks with more than 2 million registered users must publish detailed transparency reports every six months or risk sizeable fines). The first comprehensive transparency report focused on content moderation was released by YouTube in April 2018, followed days later by Facebook, which

[392] Michael Karanicolas, 'A FOIA for Facebook: Meaningful Transparency for Online Platforms' (2021), at page 7.

[393] 'The Santa Clara Principles on Transparency and Accountability in Content Moderation', https://santaclaraprinciples.org/open-consultation/ (accessed 18 January 2022).

provided more detailed explanations of its content-moderation processes and a redacted and simplified version of the guidelines provided to moderators. This was soon followed by reports from other internet companies.

These transparency reports contain aggregate data regarding government requests for user information, government demands for the removal of content alleged to violate local law, preservation requests, takedowns related to intellectual property, and aggregate data relating to information about the broad categories into which deleted content falls (as per the community guidelines) as well as aggregate data of the numbers of taken down, appealed, and restored posts. Increasingly, they also report what percentage of the material was removed before being reported on by users (what they refer to as 'proactive removal'). While most major tech companies have endorsed the principles of the Santa Clara Declaration, larger companies, such as Facebook, YouTube, and Twitter, have been reported as significantly lacking in their implementation of the principles.[394] A revised set of principles was published in 2021, which noted that, although platforms have expanded their content-moderation practices to include algorithmic tools, they have not provided sufficient transparency on how they are developed and used.[395]

While we will return to the issue of opacity, the reports by Facebook, Twitter and TikTok do provide a wealth of information that allows us to place the issue of hate speech on social media in South Africa into a broader, global context. The following pages briefly summarise some of the key aspects of the data relating to hate speech contained in the latest available transparency reports from Facebook, Twitter, and TikTok.

Hate speech moderation data on Facebook

Facebook has required moderators to label the violation in question on a piece of content since 2017 (with its automated detection technology doing the same) in order to provide more granular information. This has allowed actioned content to be disaggregated according to the broad community guideline policy it has violated. While such disaggregation is welcome, this process is only done according to the broad community guideline labels (such as violence, hate speech, nudity) and thus provides no indication of what form the hate speech in question took.[396] Facebook's transparency report relating to Hate Speech for the third quarter of 2021 (July to September 2021) suggests that 0.03% of the content posted on the platform contained hate speech.[397]

[394] Spandana Singh, 'Assessing YouTube, Facebook and Twitter's Content Takedown Policies: How Internet Platforms have Adopted the 2018 Santa Clara Principles', *Open Technology Institute*, 7 May 2019.

[395] 'The Santa Clara Principles on Transparency and Accountability in Content Moderation', https://santaclaraprinciples.org/open-consultation/ (accessed 18 January 2022).

[396] 'How we label violations' https://transparency.fb.com/en-gb/policies/improving/content-actioned-metric/ (accessed 7 February 2022).

[397] It is important to note that prevalence here has been measured as the estimated number of views that showed violating content, divided by the estimated number of total content views on Facebook.

This consisted of 22.3 million instances of content actioned, down from 31.5 million in the previous quarter (see Figure 3.10.1).[398] It is important to note that 'content' here has a very specific definition. Facebook's transparency centre notes that a post with no photo or video or a post with a single photo or video counts as one piece of content. However, if, for example, a post has multiple photos, each will count as a separate piece of content. So, if a Facebook post is removed that contains text and four photos, this would count as five pieces of content actioned. It is also, however, important to note that, if an account is removed, only the content that action was explicitly taken on is counted towards this total.[399]

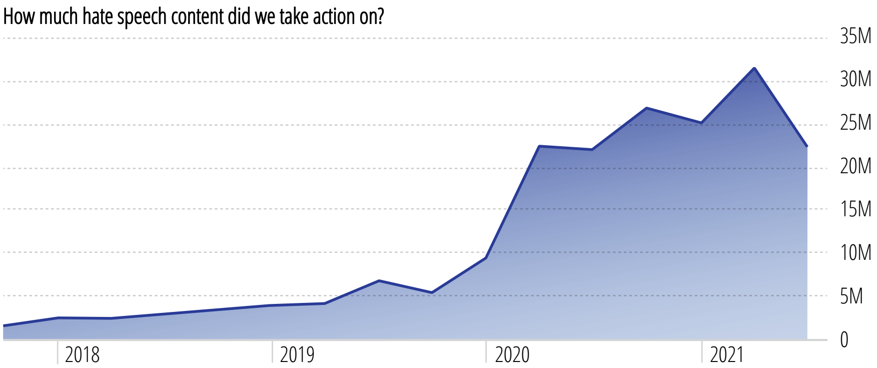


Figure 3.11.1: *Content actioned by Facebook for Hate Speech.*[400]

In response to this actioned content, there were 1.1 million appeals. In total, 394 000 pieces of actioned content were later restored (303 000 of these were restored without appeal). This suggests that only 91 000 appeals were successful, less than 10% of the total of overall appeals. This statistic suggests that users tend not to appeal content-moderation decisions relating to hate speech. There may be various reasons for this. The users might accept that they have violated community guidelines, they may not bother reporting due to the possibly limited nature of the censure, they may believe that the process is so opaque that there is little point in appealing, or users may have simply left the platform for other social-media platforms with less restrictive content-moderation practices. What these figures do suggest, however, is that there is very little debate and discussion that occurs between user and platform in relation to material removed on the basis of hate speech.

[398] https://transparency.fb.com/data/community-standards-enforcement/hate-speech/facebook/#prevalence (accessed 30 January 2022).

[399] 'How we count content and actions', https://transparency.fb.com/en-gb/policies/improving/content-actioned-metric/ (accessed 7 February 2022).

[400] 'How much hate speech content did we take action on' https://transparency.fb.com/data/community-standards-enforcement/hate-speech/facebook/ (accessed 7 February 2022).

Perhaps the most notable portion of the report relates to the 'proactive rate' figures. The proactive rate refers to the number of pieces of content acted on that were found and flagged before any Facebook users reported them divided by the total number of pieces of content on which action was taken. This has shifted from 76.4% of flags being reported by users between October and December 2017 to only 3.5% being flagged by users between July and September 2021 (see Figure 3.10.2). This suggests the increasing reliance on algorithmic methods to flag content, a process that was accelerated by the Covid-19 pandemic.

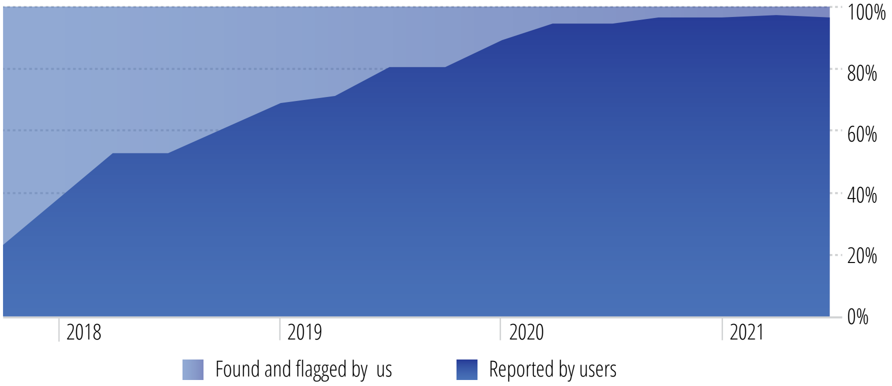


***Figure 3.11.2:** Proactive rate of content actioned by Facebook*[401]

While this shift is indicative of an increasing reliance on automated methods of removal that can delete content before it is seen by anybody else, it is important to recognise the importance of human content moderation to the effectiveness of automated techniques. Facebook's 'organic content policy' manager, Varun Reddy, highlighted this issue in a February 2021 interview when he stated that the decrease in the availability of commercial content moderators as a result of Covid-19 lockdowns in 2020 has had a significant impact on the effectiveness of the content-moderation algorithms, which are trained using the data produced by human moderators. The lack of new training data will thus have an impact on their effectiveness over time.[402]

The last piece of information that we wish to highlight is the data relating to the number of requests made for user data by the South African government (see Figure 3.10.3). Since data on this issue was first published in 2013, the number of requests has never been higher than 23 over a six-month reporting period and there have only been fifteen such requests in the first half of 2021. This suggests that the South African government seems to pay little attention to the material

[401] https://transparency.fb.com/data/community-standards-enforcement/hate-speech/facebook/ (accessed 7 February 2022).

[402] Ananya Bhattacharya, 'How Covid-19 lockdowns weakened Facebook's content moderation algorithms', *Quartz India*, 24 February 2021.

posted on Facebook or simply does not have the resources to do so. This is made starkly clear when South Africa's figures are compared to that of the United Kingdom, which made 10 678 requests in the first half of 2021 with a minimum of 6000 requests for each six-month reporting period since the second half of 2016.

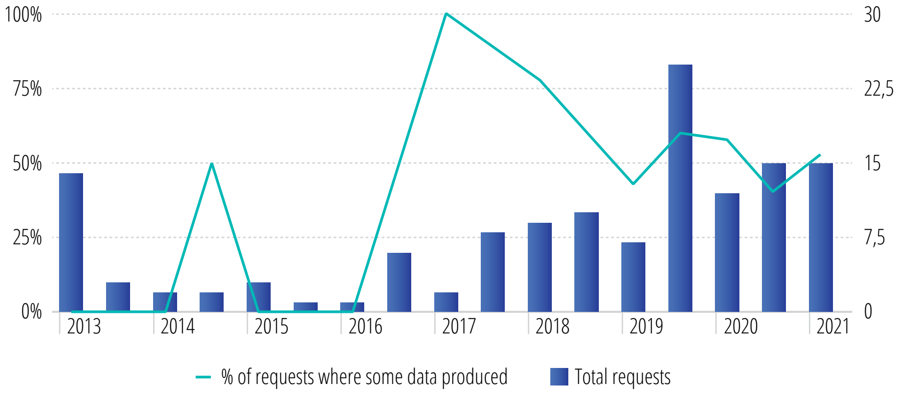


Figure 3.11.3: Government requests for user data, South Africa.[403]

Hateful conduct moderation data on Twitter

Similar trends regarding South African government requests can also be seen in Twitter's transparency reports. Since reporting into information and removals requests by governments began to be reported in 2012, the South African government has made three information requests and six legal requests (which Twitter defines as "subpoenas, court orders, or other legal documents that cite a statute or other law in association with some sort of claim or demand").[404] In comparison, the United Kingdom has made 8062 information requests and 839 legal requests. It will be interesting to see whether the number of requests to Twitter will increase following a report by the Centre for Analytics and Behavioural Change that certain Twitter accounts played a crucial role in amplifying calls for unrest, rioting, looting or violence during the civil unrest that occurred in July 2021 in South Africa.[405]

The most recent figures also suggest there was a large spike in the number of accounts actioned in the six-month period from January to June 2021. According to the latest Twitter Transparency Centre report, in the six months between January and June 2021, there were 12.9 million reported accounts (see Figure 3.10.4).

[403] https://transparency.fb.com/data/government-data-requests/country/ZA/ (accessed 7 February 2022).

[404] 'My account was named in a legal request. What does this mean?', https://help.twitter.com/en/rules-and-policies/twitter-legal-faqs (accessed 7 February 2022).

[405] 'The Dirty Dozen & the Amplification of Incendiary Content during the Outbreak of Unrest in South Africa, July 2021', Centre for Analytics and Behavioural Change (2021).

Twitter actioned 4.8 million accounts (up from just over 3.5 million accounts in the previous reporting period), suspended 1.2 million accounts (up from just over 1 million accounts) and removed 5.9 million pieces of content (up from just over 4.4 million). Twitter only began releasing content-moderation data in 2019, so it is not clear whether this increase in content-moderation numbers is simply a result of increased overall users or due to more effective content-moderation practices. The large majority of the increased amount of moderated content is made up of a category labelled 'sensitive media'. These posts are not removed but are marked as sensitive and placed behind a warning message that needs to be acknowledged before the media can be viewed, allowing those who want to avoid such sensitive content to easily do so if they wish. This form of marking up 'sensitive media' has spread to all of three social-media platforms under consideration and can be seen as a strategy of getting users to ignore material that may offend their sensibilities and therefore limit the amount of content that is flagged by users.

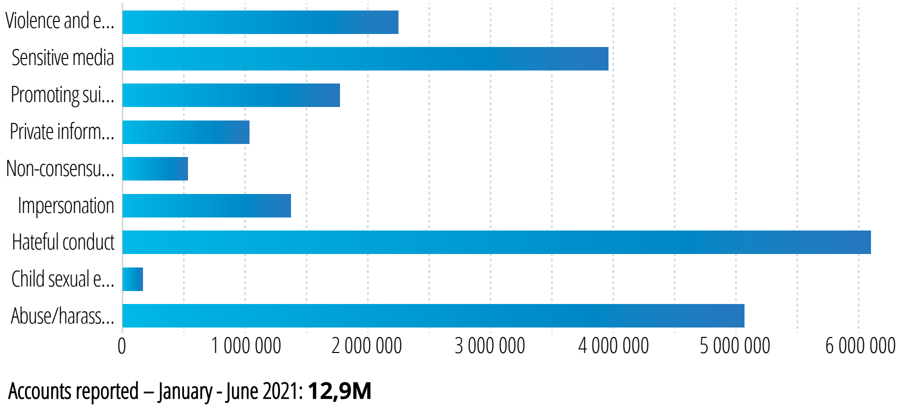


Figure 3.11.4: *Number of accounts reported to Twitter.*[406]

[406] https://transparency.twitter.com/en/reports/rules-enforcement.html#2021-jan-jun

Policy Category	Accounts actioned	Accounts suspended	Content removed
Total	**4 826 539**	**1 240 149**	**5 913 337**
Abuse / harrassment	1 043 525	99 565	1 547 654
Child sexual exploitation	456 146	452 754	6 087
Civic integrity	581	23	593
COVID-19 misleading information	27 935	617	33 761
Hateful conduct	1 108 722	133 585	1 606 979
Illegal or certain regulated goods or services	175 798	87 530	420 950
Impersonation	216 846	199 229	21 188

Figure 3.11.5: *Number of accounts actioned, accounts suspended, and content removed by Twitter per category from January to June 2021.*

Policy Category	Accounts actioned	Accounts suspended	Content removed
Total	**4 826 539**	**1 240 148**	**5 913 337**
Non-consensual nudity	29 635	7 519	64 596
Private information	30 714	3 178	54 596
Promoting suicide or self-harm	345 100	8 621	413 769
Sensitive media	1 630 554	164 260	1 655 608
Terrorism / violent extremism	44 974	44 974	0
Violence	89 245	66 445	101 907

Figure 3.11.6: *Number of accounts actioned, accounts suspended, and content removed by Twitter per category from January to June 2021 continued.*

Under the broad category of 'Hateful conduct', just over 6.1 million accounts were reported out of the overall total of 12.9 million. Of these 6.1 million accounts, 4.8 million were actioned with 1.2 million of these accounts suspended.

It would seem then that, while half of all reported accounts were reported for hateful conduct, only one-quarter of all actioned accounts were actioned for such conduct. This figure drops even more precipitously when we consider the number of accounts suspended. Of the just over 1.2 million suspended accounts, 133 585 were suspended for containing hateful content, about one-tenth of overall suspended accounts (see figures 3.10.5 and 3.10.6). Meanwhile, of the total number of content items removed (a total of 5.9 million), just over 1.6 million were hateful content. This suggests that, as discussed above, Twitter seems relatively reluctant to suspend accounts due to hateful conduct.

In its latest Transparency Report, Twitter has also added a new metric – 'impressions' – which capture the number of views a tweet received prior to removal. The webpage for this new metric was not active at the time this report was written, but the summary of the statistics given in the Transparency Report are illuminating, nonetheless. The report claims that from 1 January 2021 through to 30 June 2021, of the 4.7 million tweets removed for violation of Twitter Rules, 68% received fewer than 100 'impressions' prior to removal, with an additional 24% receiving between 100 and 1000 impressions and 8% receiving "more than 1000 impressions". The report goes on to claim that violative tweets accounted for less than 0.1% of all impressions for all tweets during that time period.[407]

As is the case with Facebook, Twitter's transparency report provides no indication of what forms this hateful conduct took, nor do they explain how these content-moderation decisions were made. In April 2019, it was reported that about 38% of abusive tweets taken down each week were being proactively detected by machine-learning models.[408] No information is easily available regarding what these numbers are today. Twitter's 2020 Transparency Report indicated that the impact of Covid-19 had led to increased use of machine learning and automation, but with no suggestion of how significant this increase has been or how it has been implemented across various categories.

Hateful behaviour moderation data on TikTok

TikTok only began providing transparency reports from January 2019 onwards for six-month periods until January 2021, since which they have switched to quarterly reports. This only occurred following threats from the Trump administration to ban TikTok in the US and following its banning in India. Its first transparency report suggested that the majority of content removed from the platform was from India (more than 16 million videos) and the US (with nearly 4.6 million).[409] Unsurprisingly, the two countries with the highest legal requests and emergency requests were the United States (with a total of 560 requests) and India (with a total of 104 requests). In comparison, South Africa had a total of one request.[410]

Its latest report claims that, from April-June 2021, just over 81.5 million videos were removed for violating its Community Guidelines or Terms of Service, less than 1% of all videos uploaded. TikTok also claims to have identified and removed 93% of these within twenty-four hours of being posted and 94.1% before a user reported them, while 87.5% were removed when they still had zero views. They also point out that just shy of 17 million of the total removals (almost 21% of the

[407] Ibid.

[408] Kalev Leetaru, 'Twitter Follows Facebook's Dystopian Path towards Unaccountable Automated Content Filtering', *Forbes*, 23 April 2019.

[409] Margaret Harding Gill, 'TikTok reveals content moderation stats amid growing global pressure', *Axios*, 9 July 2020.

[410] https://www.tiktok.com/safety/resources/transparency-report-2020-2?lang=en (accessed 7 February 2022).

total) were removed by “technology [...] that automatically detects and removes some categories of violative content“.[411]

In July 2021, the platform stated that content flagged by technology tools was also reviewed by a content moderator but that this would shift to a process of automatically removing some types of content that violated policy over minor safety, adult nudity and sexual activities, violent and graphic content, and illegal activities and regulated goods, in order to allow its commercial content moderators to focus on more contextual and nuanced areas such as hateful behaviour.[412] Just a few months later, however, the platform was fielding a wave of user complaints about content take-downs and account suspensions.[413]

When it comes to ‘Hateful behavior’ as a category (which includes hate speech and what is referred to as ‘hateful ideologies’), 2.2% of the overall videos removed violated this policy, a higher percentage of overall content than is the case with Facebook and Twitter. The report goes on to claim that 80.8% of these videos were removed within 24 hours of being posted, 60.6% were removed at zero views, and 72.9% were removed before any reports were logged – what they refer to as ‘proactive removal’ (see Figure 3.10.7).

The use of the term ‘proactive’ here should also be treated wearily. TikTok, as with all of these social-media platforms that form part of this study, practice a ‘post first monitor later’ policy. So it is not proactive in the sense of preventing posts that violate the community guidelines from making it on the platform to begin with. The ‘hateful behaviour’ category, along with that of ‘Harassment and bullying’, had significantly lower proactive removal rates and overall removal rates compared to the other categories, something that is a feature across each of the platforms. This highlights the difficulty of algorithmic removal of such content, but also suggests that hateful content stays on these social-media platforms for far longer than any other form of violation.

[411] https://www.tiktok.com/transparency/en/community-guidelines-enforcement-2021-2/

[412] Tiyashi Datta, ‘TikTok to automatically remove content that violates policy’, *Reuters*, 9 July 2021.

[413] Karl Herchenroeder, ‘TikTok Sees Wave of Takedown Disputes after Automation Shift’, *Communications Daily*, 11 August 2021, https://communicationsdaily.com/news/2021/08/11/TikTok-Sees-Wave-of-Takedown-Disputes-After-Automation-Shift-2108100053#:~:text=TikTok%20U.S.%20Safety%20head%20Eric,block%20or%20delete%20any%20posting (accessed 8 February 2022).

	Proactive removal rate	Removal within 24 hours rate	Removal at zero views
Adult nudity and sexual activities	90.3%	90.0%	78.5%
Harassment and bullying	73.3%	83.8%	61.4%
Hateful behavior	72.9%	80.8%	60.6%
Illegal activities and regulated goods	97.1%	95.7%	92.3%
Integrity and authenticity	88.3%	86.2%	67.9%
Minor safety	97.6%	95.4%	93.9%
Suiside, self-harm and dangerous acts	94.2%	90.8%	81.8%
Violent and graphic content	94.9%	94.3%	86.6%
Violent extremism	89.4%	90.1%	79.5%

Figure 3.11.7: *Removal rate of videos removed for violating TikTok's Community Guidelines or Terms of Service from April to June 2021.*

The unstated problem of false negatives and false positives

While the amount of content removed may sound impressive, given the scale of social-media activity, substantial numbers of posts containing hate speech remain on these platforms. These figures also ignore the fact that often substantial numbers of people will see such a post before it is taken down and that material can spread rapidly within the same platform or across various platforms. It is also important to note the general pivot in the social media industry as a whole towards automatic content moderation as a panacea for online harm. Recent research indicates how ineffective automated content moderation can be.

The effectiveness of human commercial content moderation has also been called into question. A 2020 report by the Centre for Countering Digital Hate found that 84% of the 714 posts they had reported containing anti-Jewish hatred across various platforms (which were viewed at least 7.3 million times) were not acted upon by social-media companies. Facebook performed the worst, failing to act on 89% of the posts, while Twitter allowed a range of hashtags used for antisemitic content (such as #rothschild, #fakejews and #killthejews).[414] These transparency figures, in short, cannot tell us the amount of potentially rule-breaking content which goes unnoticed (ie, false negatives), 'awful but lawful' content that is removed, or the total amount of innocent content that is wrongly flagged (false positives).[415]

[414] Center for Countering Digital Hate, Failure to Protect, report available at https://www.counterhate.com/failuretoprotect.

[415] Michael Karanicolas, 'A FOIA for Facebook: Meaningful Transparency for Online Platforms' (2021), at page 4.

For example, Maarten Sap et al point out how terms previously used to disparage communities (for example, “n*gga“, “queer“) have been reclaimed by those communities and are therefore unoffensive when used among these groups, while remaining offensive when used by outsiders. However, annotators’ insensitivity to such differences can lead to racial bias both in human content moderation and in the automatic hate speech detection models (or algorithmic models) that their decisions are used to train (see Figure 3.10.8). According to Sap et al, even though hate speech often targets minority groups like LGBTQ+ communities and African Americans in the US, models that have been trained on such datasets have in fact not only acquired but also propagated these biases. As a result, tweets by self-identified African Americans using African American English are up to two times more likely to be labelled as offensive.[416]

Given the low rates of appeal against the removal of content and banning of accounts on social-media platforms, it seems likely that a great deal of such innocent content is caught in the content-moderation trawl, providing good reason for various groups to believe that they are being systematically targeted by content-moderation practices. These beliefs are likely to increase with the concerted move towards automated content removal, and to come to the fore even in cases where there may have been legitimate reasons for censure, which may remain unknown to the user due to the opacity of the content-moderation process highlighted earlier in this report.

[416] Maarten Sap et al, ‘The Risk of Racial Bias in Hate Speech Detection’ in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (2019), pp. 1668–1678. For another study that came to similar conclusions, see Thomas Davidson, Debasmita Bhattacharya & Ingmar Weber, ‘Racial Bias in Hate Speech and Abusive Language Detection Datasets’ in *arXiv preprint arXiv:1905.12516* (29 May 2019).

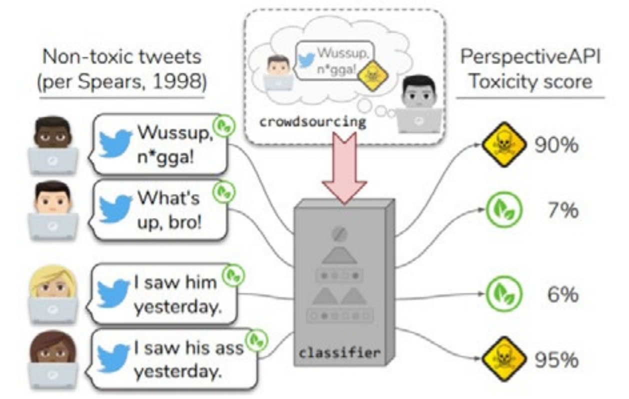


Figure 1: Phrases in African American English (AAE), their non-AAE equivalents (from Spears, 1998), and toxicity scores from PerspectiveAPI.com. Perspective is a tool from Jigsaw/Alphabet that uses a convolutional neural network to detect toxic language, trained on crowdsourced data where annotators were asked to label the toxicity of text without metadata.

Figure 3.11.8: *A graphic representation of the risk of racial bias in hate speech detection.*[417]

The transparency illusion

While these transparency reports are certainly a step in the right direction as they at least provide us with an idea of what the tip of the iceberg looks like when it comes to hate speech and other community guidelines violations, Svea Windwehr and Jillian C. York point out that "transparency does not always equal transparency". Focusing on Facebook's August 2020 transparency report, they note that the report itself is emblematic of some of the deficits of companies reporting on their own content-moderation practices. More importantly, they point out that content moderation and its impact are always contextual, and the sterile reportage of numbers and percentages does not tell us why or how these decisions are taken. They suggest that transparency is a misnomer here:

> *Actual transparency should allow outsiders to see and understand what actions are performed, and why. Meaningful transparency inherently implies openness and accountability, and cannot be satisfied by simply counting takedowns. That is to say that there is*

[417] Maarten Sap et al, 'The Risk of Racial Bias in Hate Speech Detection' in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (2019), pp. 1668–1678, at page 1668.

> ***a difference between corporately sanctioned 'transparency,' which is inherently limited, and meaningful transparency that empowers users to understand Facebook's actions and hold the company accountable.***[418]

One of the main critiques of these reports is the fact that they simply provide aggregate statistics. For example, there is no sense of data by country or what particular forms of hate speech or hateful conduct took place, nor is there an indication of which types of conduct were most likely to lead to censure. There is also very little sense of what actions were actually taken. Was the removal a partial or full removal? Was it a country-specific or global removal? Where does shadow-banning fit in these statistics? There are also no suggestions as to how the definitions for hate speech are operationalised. More importantly, given the increased use of AI tools during the Covid-19 pandemic and the well-documented shortcomings of AI tools to judge the social, cultural, and political context of speech correctly (as seen in the example above), there is no indication of what materials both human and machine automated tools are trained on or of the relationship between, and oversight of, these forms of review.[419]

These aggregated statistics make it difficult for us to draw conclusions about the quality of the decisions and what this quality consists of across the interrelated content-moderation processes.[420] In addition to this, each platform has its own policy regulating what it deems acceptable and while, as we saw above, there is a great deal of overlap in the wording of these policies and the categories listed, the extent of flagging on each platform may vary considerably depending on its user base, the automated moderation tools used, and what data these tools are trained with. In addition, the directives given to commercial content moderators and the manner in which they moderate may differ significantly.[421]

In many ways, the transparency reports actually make the practices of these platforms even more opaque rather than transparent. More importantly, this information does not actually allow us to identify how well these moderation systems are working or how they can be improved.[422] The belief in the effectiveness of transparency reports, Mike Annany and Kate Crawford note, is built on the belief that, the more information is made available, the more defensibly an institution can be governed and held accountable and that this performative act will help to produce understanding. They instead argue that, in the case of the transparency

[418] Svea Windwehr and Jillian York, 'Thank You For Your Transparency Report, Here's Everything That's Missing', *Electronic Frontier Foundation*, 13 October 2020.

[419] Svea Windwehr and Jillian York, 'Facebook's Most Recent Transparency Report Demonstrates the Pitfalls of Automated Content Moderation', *Electronic Frontier Foundation*, 8 October 2020.

[420] Nicolas P Suzor, Sarah Myers West, Andrew Quodling & Jillian York, 'What Do We Mean when We Talk about Transparency toward Meaningful Transparency in Commercial Content Moderation?' in *International Journal of Communication,* Vol. 13 (2019), pp. 1526–1543, at page 1538.

[421] GK Young, 'How much is too much: The difficulties of social media content moderation' in *Information & Communications Technology Law* (2021), pp. 1–16, at page 4.

[422] Nicolas P Suzor, Sarah Myers West, Andrew Quodling & Jillian York, 'What Do We Mean when We Talk about Transparency toward Meaningful Transparency in Commercial Content Moderation' in *International Journal of Communication,* Vol. 13 (2019), pp. 1526–1543, at page 1528.

reports of social-media platforms, this simply creates a 'transparency illusion'.[423] They provide detailed reasons why transparency cannot promise consequential accountability which, for our purposes, can be summarised under these key points:

- If transparency does not lead to any meaningful effects, then it loses its purpose, particularly if systems are not in place to create change. If a practice continues after it has been made transparent and critiqued, it simply leads to increased cynicism. In short, transparency does not necessarily build trust across different stakeholders.
- Transparency can lead to opacity when too much information leads to important information being hidden. This sometimes occurs inadvertently, but just as often is a strategic choice to distract and conceal information.
- Transparency can privilege seeing information over understanding it.
- The changing scale of content moderation, shifts in platform interfaces to flag this material, constantly changing community guidelines, and differences in how and why content moderation takes place across platforms means that there is very little standardisation in the information presented even across the transparency reports of an individual platform. It is often more productive to see what is changing across reports and how it is changing to gain insights into moderation processes across social-media platforms.
- Transparency can sometimes be harmful (for example, releasing details of flaggers may lead to retaliation in the online and offline worlds, which may expose vulnerable groups to intimidation.

For Crawford and Annany, the questions we should ask ourselves are what is being looked at, what good comes from seeing it, and what are we *not* able to see? The simple production of transparency reports is meaningless unless it is clearer what exactly is being held to account. Is it the companies themselves, the content-moderation process, the algorithmic processes used, the interface, the nature of geoblocked content, the problematic content that remains, or the creation of standardised data? Each of these questions in fact requires a very different set of practices and data. There is also the question of who exactly these companies would be accountable to. Is it accountable to its users, its shareholders, government (if so, which government), etc?

This boils down to two key issues: what are these companies accountable for and who are they accountable to? These aggregate statistics may provide an overview of content-moderation processes and broad areas of concern but are not sufficient to enable the detailed analysis necessary to hold platforms accountable.[424] The transparency reports are a process of self-reporting, which, while clearly a step in the right direction, are being done according to each company's own standards and

[423] Mike Annany and Kate Crawford, 'Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability' in *New Media & Society*, Vol. 20, No. 3 (2018), pp. 973–989.

[424] Nicolas P Suzor, Sarah Myers West, Andrew Quodling & Jillian York, 'What Do We Mean when We Talk about Transparency toward Meaningful Transparency in Commercial Content Moderation' in *International Journal of Communication,* Vol. 13 (2019), pp. 1526–1543, at page 1529.

definitions, none of which were democratically imposed. In addition, they seem to also be deployed to ward off liability claims and increased regulation rather than to facilitate greater accountability.

This point is highlighted by *The Guardian*'s 'Facebook Loophole' series, which showed how Facebook allowed major abuses of its platform in poor, small and non-western countries to prioritise abuses that attract media attention or affect the US and other wealthy countries. Sophie Zhang, a former data scientist at Facebook who worked within the company's "integrity" organisation to combat inauthentic behaviour, claimed that, "There is a lot of harm being done on Facebook that is not being responded to because it is not considered enough of a PR risk to Facebook [...] The cost isn't borne by Facebook. It's borne by the broader world as a whole".[425]

While having government oversight of the process in each country may offer some kind of solution in some places, these nations are often those with the least resources to spare for such processes. Also, in some of these nations, it may be the government itself that is spreading hate speech. The ability of third parties, such as civil society groups, to play this oversight role is heightened in such spaces, as is their need to access accurate and comprehensive information about how the content-moderation systems of social-media platforms are functioning in order to provide meaningful feedback on their efficacy.[426] Michael Karanicolas notes that this is even more difficult in the developing world, "where researchers [...] face a constant struggle to find accurate data regarding how policies are being implemented, or even which policies are operative in a particular region".[427] This is a critique that we can vouch for, given our own experiences in conducting research for this report.

A further series of reports by *The Wall Street Journal,* labelled 'The Facebook Files', based on a series of reviews of internal Facebook documents, research reports, online employee discussions and drafts of presentations to senior management, highlighted the company's research into the possible negative impacts on the day-to-day lives of many of its users. The series concluded that Facebook "knows, in acute detail, that its platforms are riddled with flaws that cause harm, often in ways only the company fully understands". It went on to conclude that:

> ***Time and again, the documents show, Facebook's researchers have identified the platform's ill effects. Time and again, despite congressional hearings, its own pledges and numerous media exposés, the company didn't fix them. The documents offer perhaps the clearest picture thus far of how broadly Facebook's problems are known inside the company, up to the chief executive himself.***[428]

[425] Julia Carrie Wong, 'Revealed: The Facebook loophole that lets world leaders deceive and harass their citizens', *The Guardian*, 12 April 2021. The theme of the costs of Facebook's actions being borne by anyone but Facebook themselves is also a theme of the more recent 'Facebook Files'.

[426] Michael Karanicolas, 'A FOIA for Facebook: Meaningful Transparency for Online Platforms' (2021), at page 9.

[427] Ibid., at page 11.

[428] https://www.wsj.com/articles/the-facebook-files-11631713039

The transparency reports hide key issues and instead aim to simply highlight how much content is being removed. They are performative rather than providing the basis for meaningful analysis and accountability for the various features that form part of the content-moderation process. For our purposes, there is no sense of what the content-moderation process for posts made in South Africa or by South Africans consists of, no meaningful data to show the extent of moderation that is occurring geographically, or who is producing the majority of the flags for this material. Is it other users who may not understand the South African context, is it South Africans who may have their own particular biases, or is it algorithmic processes?

For Windwehr and York, meaningful transparency requires the clarification of the number of human moderators available, the training and guidelines received, the languages covered, whether there are no native language speakers for particular languages, the ratio of moderators per language, more light being shed on the algorithmic content-moderation systems, the inputs that are used for these systems, whether their purpose is simply to flag content or also to judge and categorise the content, how many of the complaints are reviewed by humans, and the relationship between human and automated reviews.

Karanicolas points out that such transparency should extend beyond the reports, as even the internal governance structures of these platforms are designed, in many cases, with an eye to curtailing meaningful oversight. For example, at Facebook, the CEO and Chair of the Board are the same person, effectively making that person accountable to themselves, which could be seen as problematic given the enormous power and public importance of a platform like Facebook.[429] Karanicolas also points out how social-media platforms reflexively use non-disclosure agreements. These extend to their content moderators, which a pending lawsuit against Facebook claims are not limited to user data and help perpetuate a culture of "excessive secrecy". The net result of this, Karanicolas argues, is that "for all transparency, the platforms have managed to insulate themselves from meaningful independent oversight 'through code and through contract'".[430] For Karanicolas, transparency requires more than relying on "the platforms' largesse in delivering scraps of information", but requires an approach to transparency based on best practices from the public sector and that moderation structures should be "open by default".[431]

[429] Michael Karanicolas, 'A FOIA for Facebook: Meaningful Transparency for Online Platforms' (2021), at page 11.

[430] Ibid.

[431] Ibid.

4 Methodology

Our approach involved coding datasets extracted via Communalytic and CrowdTangle, as well as digital ethnography, characterised by the researcher “following“ the medium.[432] The material was coded by eight annotators before being reviewed by academics based at the Kaplan Centre for Jewish Studies, the Department of Political Studies, and the Department of Historical Studies at the University of Cape Town.

Annotators

A diverse group of eight individuals proficient in several of South Africa’s national languages were hired and trained to annotate extracted social-media content. Annotation entailed flagging posts considered to contain racial discrimination, xenophobia, and/or antisemitism, as well as posts that were considered to provide a sense of the broader discourse around these issues.

As part of their training, the eight members of the annotation team participated in introductory sessions focused on the history of racism and antisemitism with a particular focus on the South African context, led by two experts in these fields. Team members were also introduced to the key documents that we used to define antisemitism, racial discrimination, and hate speech: the Jerusalem Declaration on Antisemitism, the Promotion of Equality and Prevention of Unfair Discrimination Act (PEPUDA or the Equality Act, Act No. 4 of 2000), and the interpretations of PEPUDA put forward in South Africa’s National Action Plan to Combat Racism, Racial Discrimination, Xenophobia and Related Intolerance, which was launched on 25 March 2019.

As has been described, the annotation team were provided with data relating to several ‘flashpoints’ from Twitter and Facebook. These flashpoints were closely tied to South African events and were chosen because they left sizeable traces on Twitter and Facebook and were widely reported on in the popular media, thus permeating South African new and old media in various ways. We then analysed material on TikTok relating to these same flashpoints.

[432] Alessandro Caliandro, ‘Ethnography in digital spaces: Ethnography of virtual worlds, netnography, & digital ethnography’ in Rita M Denny and Patricia L Sunderland (eds.), *Handbook of anthropology in business* (Abongdon: Routledge, 2016), pp. 658-679.

Software

This project deliberately followed a 'low tech' approach to data extraction and annotation. A program called Communalytic was used to extract data from Twitter. Communalytic, developed by the Social Media Lab at Ryerson University in Toronto, is an accessible research tool for studying online communities and discourse. Communalytic is a search-based tool, making the subject matter more accessible to researchers in the social sciences. Communalytic was used in tandem with Twitter's newly introduced Academic Research API, which allows tools such as Communalytic to retrieve current and historical data from Twitter.[433] Data pulled from Twitter includes the 'tweet id', text of the tweet, username, and number of retweets, replies, likes and followers. Facebook data was collected from CrowdTangle, a tool developed by Facebook that can track and collect public posts. Data pulled from Facebook includes the page and username, country of page admin, text of post, and number of likes, comments, shares and interactions. The data from both platforms is pulled into a spreadsheet that can then be used for further analysis.

Datasets

The first 'flashpoint' flagged for analysis was a series of protests in the Cape Town suburb of Brackenfell. Social media data relating to the event were extracted from Twitter and Facebook using the search terms "#EFFinBrackenfell" OR "EFF Brackenfell" (delimited from October 2020 to May 2021). This produced 68 746 tweets. This was reduced to 10 892 tweets following the removal of all retweets (57 854) from the dataset. A random sample of 10% of the remaining original tweets was extracted to produce a final dataset of 1089 tweets. The same search terms returned 6333 posts from Facebook. The dataset also indicated the number of comments per post – with a total of 462 027 comments across all posts. While needing to reduce the size of the dataset, we also wanted to focus on posts with the greatest number of comments. As such, we dropped all posts with zero comments (leaving 3781 posts). We then dropped posts where the number of comments was below the 80th percentile (this meant dropping posts with less than 101 comments). This left 765 posts remaining, of which we selected a random sample of 500 posts for annotation.

The second flashpoint that was analysed involved the protests that erupted in the small town of Senekal. Material relating to the Senekal protests was extracted using the terms "Senekal EFF" or "Brendin Horner" or "Senekal protest" (for the period October 2020 to May 2021). This search produced 65 488 tweets. This was reduced to 10 605 tweets with the removal of the 54 883 retweets from the dataset. A random sample of 10% of these remaining tweets was taken to produce a final dataset consisting of 1061 tweets. The search also returned 9565 Facebook posts (with 444 349 comments across all posts). We dropped all posts with zero comments (leaving 5629 posts) and similarly excluded posts where the number

[433] API refers to *application programming interface*. The API is essentially a software intermediary that allows two applications to 'speak to one another', acting as an as interface between programs.

of comments was below the 80th percentile (dropping posts with less than 66 comments). This left 1119 posts, of which a random sample of 500 posts was extracted for annotation.

The third flashpoint was that of Operation Dudula. We searched Twitter and Facebook (for the period January 2020 – July 2021) using the hashtags #operationdudula and #dudula. This search produced a smaller dataset of 3779 tweets, with a final dataset of 581 tweets once the 3198 retweets were dropped from the sample. Given the size of the dataset, all 581 tweets were coded. The search returned 135 Facebook posts (with 2423 comments across all posts). The dataset was similarly analysed in its entirety.

The final flashpoint focused on the conflict in Gaza in May 2021. We searched Twitter and Facebook using the search terms "Israel" or "Gaza" or "Palestine". The search period was the month of May 2021. As these search terms are not specific to a local/South African context, they yielded an overwhelming amount of content. As such, we only extracted tweets from users with their profile country set to South Africa on Twitter and, similarly, extracted Facebook posts where the country of the page admin was specified as South Africa. The search yielded a dataset of 2440 tweets (with the search method, Communalytic only provided original tweets and no retweets). A random sample of 1000 tweets was analysed. Additionally, the search yielded 2964 Facebook posts, with a total of 338 251 comments. Here we focused mainly on pages of local media institutions.

Ethnography and Annotation

When the final datasets were extracted for coding, the eight team members were split into two groups. Four continued with an ethnographical approach that involved searching Facebook pages for the most useful publicly available material relating to the chosen flashpoints. The remaining four began to annotate the extracted datasets. These four annotators were in turn split into two pairs, with each pair focused on a different flashpoint. These datasets would then be compared and contrasted by another member of the team to measure 'intercoder reliability', which refers to the extent to which independent coders reach the same conclusion after evaluating the same post. The process of checking intercoder reliability highlights cases of complete agreement between annotators, as well as reveals the variability between the interpretations of coders who have received identical training.

Using the 'tweet id' – a unique 18-digit number given to each post that can be used to access it – the annotators opened a web page that showed the tweet as part of the original thread in which it was posted. After reading the relevant tweet and the broader thread in which it appeared, as well as visits to the home page of the tweeter in question, the annotators then coded the relevant fields. After taking a screenshot of the tweet in question to ensure we would still have access to the material even if it was later removed by the user or by Twitter itself, annotators captured the language used within the tweet and decided whether the extracted tweet constituted racial discrimination, antisemitism, and/or xenophobia. If they felt it constituted hate speech, they listed the sections of the Jerusalem Declaration and/or PEPUDA that they believed it contravened.

Annotators also provided a sentiment rating for the tweet (this involves returning a score to measure how positive or negative a post is), determined whether it involved the “calling out“ of another user, was sarcastic in nature, and if it contained terms useful in building a lexicon of hate speech terms in the South African context. This was all done drawing on the diverse language skills of the various members of the team.

The coded datasets were analysed by the core research team of Dr Thierry Rousset, Dr Gavaza Maluleke, and Prof Adam Mendelsohn, with assistance from Dr Kerri Serman (quantitative analysis), Dr Ethan Roberts (quantitative analysis), and Patricia Chirwa (network analysis).

ISBN: 978-0-6397-3015-8 (e-book)

Kaplan Centre
University of Cape Town
Rondebosch 7701
Cape Town
Tel: +27 (0)21 650 3062
www.kaplancentre.uct.ac.za

Johannesburg Holocaust
& Genocide Centre
1 Duncombe Road
Forest Town 2193
Johannesburg
Tel: +27 (0)11 640-3100
www.jhbholocaust.co.za

Konrad Adenauer
Stiftung
60 Hume Road
Dunkeld 2196
Johannesburg
Tel: +27 (0)11 214 2900
www.kas.de/mediaafrica