Reporting of online harms and threats against the Covid-19 Response: Fortnightly Update

SENSITIVE
11 July 2022
For the period 27 June to 7 July 2022

Kate Hannah, Sanjana Hattotuwa, Kayli Taylor
The Disinformation Project (TDP)

For further inquiries or should this report fall within the scope of an Official information Act request, contact the Disinformation project Director, Kate Hannah

s9(2)(a)

Executive Summary

- The ever-changing mis- and disinformation landscape continues to evolve and shift, with impacts on democracy, stable governance, and human lives.
- Social media groups and channels studied have subscriber numbers in the hundreds of thousands: 380,000 subscribers across 161 channels on Telegram; 956,685 followers across 95 Facebook pages; 220,442 followers over 47 Instagram accounts.
- The Ministry of Health's communications are providing another tool for mis- and disinformation producers to use to stir up fear and mistrust in public health institutions and responses.
- The overturning of Roe v Wade and unveiling of anti-abortion sentiment could have an effect on the safety of abortion providers (both individuals and infrastructure), and to those seeking abortion services in Aotearoa New Zealand.
- QAnon rhetoric, further propelled by the recent re-emergence of Q, is present, embedded, and growing in the ecologies we study.
- s9(2)(c)
- Mis- and disinformation subscribers rely on racism to blame and ostracize. This has implications for social cohesion and contributes to a muddying of conversations about 'free speech'.
- Frames and themes of misogyny dominate ecologies we study – with effects on norm-shifting and changing what is 'acceptable' in Aotearoa New Zealand.
- Queerphobia and harms against the LGBTQ+ community are commonplace, with religious and nationalist frames being used to scapegoat.
- Sov-Cit rhetoric and the use of 'paper terrorism' are increasing in Aotearoa New Zealand – which will have effects on frontline Police, and the way Justice and other agencies operate.
- The genuine cost of living crisis and fuel increases are having an impact on the way New Zealanders feel and live – something that is being weaponised by mis- and disinformation producers.
- The belief that the Christchurch Terror Attack was a 'false flag' continue to grow in prominence in the ecologies we study, with effects on whānau of victims, the Muslim community, mistrust in government, and de-sensitisation to violence.

Introduction

Disinformation is a threat to democracy,[1] stable governance,[2] and human life.[3] Since the start of the Covid19 pandemic and its associated infodemic,[4] disinformation and its impact on people and society in Aotearoa New Zealand has grown. The Disinformation Project (TDP) has analysed this since February 2020, paying particular attention to the volume, velocity, and vector of information. We use daily data collection and analysis (which form the basis of this summary report), use computational and manual tools to scan open-source social media post and commentary across a wide range of social media platforms, websites, and media/alternative media organisations. The information landscape studied is developed using 'snowball' techniques, which means that we have expanded the inclusion of pages, groups, and channels only when they are signalled by existing locations of study.

Within the social media ecologies studied, key individuals and groups producing mis- and disinformation capitalise on growing uncertainty and anxiety amongst communities, related to Covid-19 public health interventions, including vaccination and lockdowns, to build fear, disenfranchisement, and division. Mis- and disinformation is also particularly targeting and scapegoating already marginalised or vulnerable communities – for whom distrust of the state is the result of intergenerational trauma and lived experience of discrimination or harm, which can increase engagement with conspiratorial explanations and disinformation. Over the past two and a half years of research, TDP has developed a thorough and balanced understanding of the harms that mis- and disinformation and 'dangerous speech' present to social cohesion, freedom of expression, inclusion, and safety. (See Appendix One for our definitions of these terms).

The landscape studied – Covid-19

The landscape studied originated as locations engaged in mis- and disinformation related to Covid-19 and the Covid-19 response, and our study, the type of content produced and shared within this landscape has shifted over time, so other narratives and themes within this landscape now form part of our analysis.

For example, on September 26, 2021, the Telegram channels we studied totalled 44,267 subscribers; as of July 1, 2022, we analyse daily 161 Telegram channels with 380,000 subscribers. While there is no feasible ethical method for de-duplication, the growth – and similar growth of both locations (pages, groups, accounts) and followers on Facebook and Instagram – signals increasing interest in these ideas, and continued engagement with content despite the shifts in narrative and theme we note above.

For the purposes of this reporting, we focus on online harms and threats against the Covid-19 response, including people and places associated with the Covid-19 response. This includes covid denialism, covid minimisation, anti-vaccination messaging (which is increasingly spilling

---

[1] https://www.europarl.europa.eu/RegData/etudes/STUD/2021/653635/EXPO_STU(2021)653635_EN.pdf

[2] Ibid.

[3] https://www.scientificamerican.com/article/covid-misinformation-is-killing-people1/; https://www.axios.com/2022/04/21/barack-obama-disinformation-social-media

[4] https://www.who.int/health-topics/infodemic

from vaccination against Covid-19 into other types of vaccination, including vaccination for tamariki), and anti-mandate/anti-public health measures messaging. These are the four dominant types of messaging that TDP observes in its study of the mis- and disinformation ecologies within Aotearoa New Zealand.

The landscape studied – shifts and developments
As the pandemic has, and continues to, shift and adapt; so too do the focuses of mis- and disinformation producers and their subscribers. We outline these more thoroughly in our two most recent public reports.[5] A short summary below.

- In our paper published November 2021, we warned about the way Covid-19 related mis- and disinformation were being used as a kind of Trojan Horse to push followers and subscribers towards far-right and extremist ideologies.
- The Parliament Protest saw a large and ideologically diverse group of people brought together to advocate for highly divergent causes. Protestors had highly divergent understandings of the protest, its intentions, and its reception within non-protestors. There is also a high chance that protestors were radicalised during the protest. For example, they may have gone to protest one issue, e.g. vaccine mandates; and instead find themselves exposed to a wide array of extremist ideology.
- The causes advocated for and against by mis- and disinformation producers is constantly shifting. Current concerns include Three Waters reform, abortion law (in the wake of the overturning of Roe v Wade in the US), the rise of Sovereign Citizen rhetoric, support for the invasion of Ukraine, and the rights of LGBTQ+ people in Aotearoa New Zealand. As the media and social landscape of Aotearoa New Zealand continues to pivot and highlight diverse issues, so too do mis- and disinformation producers and their subscribers adopt new concerns.
- The end of the Parliament Protest does not symbolise a neat ending to conspiratorial thought in Aotearoa New Zealand. Such ideologies continue, adapting and growing at pace. The responsibility falls to renewing our efforts for social cohesion, honouring Te Tiriti o Waitangi, and reflecting critically on our past, shared present, and collective hopes for the future.

---

[5] Hannah, Kate, Sanjana Hattotuwa, and Kayli Taylor. "Mis- and Disinformation in Aotearoa New Zealand from 17 August to 5 November 2021." 2021.: https://thedisinfoproject.org/wp-content/uploads/2022/04/2021-11-09-FINAL-working-paper-disinformation..pdf; Hannah, Kate, Sanjana Hattotuwa, and Kayli Taylor. "The Murmuration of Information Disorders: Aotearoa New Zealand's Mis- and Disinformation Ecologies and the Parliament Protest." 2022.: https://thedisinfoproject.org/wp-content/uploads/2022/05/The-murmuration-of-information-disorders-May-2022-Report-FULL-VERSION.pdf

Health disinformation

Health disinformation, and the responses to Covid-19 and other disease, form the bulk of the focus on mis- and disinformation producers within Aotearoa New Zealand – and thus TDP's research focus.

One narrative that proliferates within the ecologies studied by TDP is that Covid-19 isn't as bad as the media, public health officials, and Governments say. This narrative emerged in 2020 and grew over 2021. In Aotearoa New Zealand, the Delta outbreak from August 2021 and onset of Omicron have fueled this narrative. This leads to the undermining of public health messaging from MoH and UAC in attempts to get (1) children vaccinated against Covid-19, (2) getting people, especially children, vaccinated against the flu, (3) getting adults boosted, and (4) getting eligible adults boosted again.[6]

Public health commentary is consistently misinterpreted. For example, an interview from Professor Michael Baker on increased mortality is repackaged by alternative media organization Counterspin Media blaming harm related to the vaccine. This conceptualization is normative in the landscape studied.

Against the backdrop of rising case numbers and the increase of deaths across the country, Counterspin Media's decontextualization is dangerous and completely at odds with the thrust of the interview. This is sophisticated disinformation – using professional media productions in service of Counterspin Media's conspiratorial and disinformation narrative production. Such disinformation production has consequences for how people interact with, and trust, mainstream media, and public health messaging.

s9(2)(g)(i)

[7] The wilful misinterpretation of this messaging is promoted by the most influential producers, such as Voices for Freedom, with widespread social media and direct marketing reach to over 200,000 New Zealanders, and likely impact on further vaccination campaigns for childhood vaccines.

Newshub Journalist Michael Morrar's piece on the poor estimation on testing capacity in the early Omicron peak,[8] and the claims of one union representative that the Ministry's stance amounted to 'misinformation' was flagged in the ecologies we study. This labelling is now instrumentalised in a number of ways: (1) to deem the term meaningless, (2) that the government was promoting misinformation (with negative effects on public trust), (3) and that

---

[6] For an example of this, see the section on harassment of Dr Jin Russell.

[7] https://www.rnz.co.nz/news/national/470282/significant-second-wave-of-omicron-may-already-be-here

[8] https://www.newshub.co.nz/home/new-zealand/2022/06/omicron-testing-backlog-review-finds-ministry-of-health-failed-to-accurately-estimate-nz-s-lab-capacity.html

research on mis- and disinformation (such as TDP) doesn't look at such 'misinformation' because of hidden agendas to silence mis- and disinformation producers and/or to silence truth.

Trust in institutions is important, particularly during a pandemic. We are observing across the ecologies we study *and* from groups with previously high trust, that trust in Government, public health institutions, and the Ministry of Health is waning. Continuation of the current erosion of trust will have dangerous implications for the way the Ministry of Health and other public-health related government organisations in Aotearoa New Zealand interact with its peoples.

International contexts

Roe vs Wade

Following the overturning of Roe vs Wade and the removal of constitutional right to abortion for people with uteruses in the United States, mis- and disinformation ecologies studied by TDP celebrated the decision, and pivoted attention to abortion laws in Aotearoa. This attitude is widespread and consistent across the diverse groups and individuals represented in our location of study. For example, a video posted to YouTube, critiquing the PM's response[9] to SCOTUS's decision by a prominent Christchurch disinformation producer has been viewed, at time of writing, over 1700 times with 240 likes and over 100 comments. The video uses graphic and inaccurate frames designed to antagonize and incite.

s9(2)(c)

---

[9] PM Jacinda Ardern slams US abortion ruling, but Chris Luxon avoids reaction, https://www.stuff.co.nz/national/129080107/pm-jacinda-ardern-slams-us-abortion-ruling-but-chris-luxon-avoids-reaction

s9(2)(c)

QAnon

QAnon is a wide-ranging and baseless internet conspiracy with origins in the United States and global influence.[11] Broadly, adherents to the conspiracy theory believe that a collection of Satan-worshipping political leaders, celebrities and billionaires rule the world – including engaging in paedophilia, human trafficking, and the harvesting of blood from children. Supporters of QAnon were involved in the attempting coup at the US Capitol building on 6 January 2021.[12]

QAnon ideas are common in the mis- and disinformation ecologies studied. The return of anonymous poster 'Q', instigator of the QAnon conspiracy theory the same weekend as the overturning of Roe v Wade saw an increase in the discussion of QAnon discourses here. One local disinformation producer posted celebrated the overturning of Roe v Wade, claiming that abortion is one way the organ trade operated in the US.

The presence of QAnon ideas is of concern in Aotearoa New Zealand for several reasons:
- QAnon, and the sentiment it encourages represents a militant and anti-establishment ideology and is associated with declining trust in institutions in the United States.
- Radicalisation, the erosion of trust in social and democratic institutions, and reductions in social cohesion are some of the effects we may see as a result of the spread of QAnon related ideologies and ideals across Aotearoa New Zealand.
- QAnon frames align closely with Russian disinformation, queerphobia, misogyny; as well as content that goes against democracy. Such content is aimed at public institutions in the United States, but is toxic to domestic cultures, communities, and contexts.

QAnon is not just limited to social media (Telegram, but present across Meta and Twitter), but takes place in offline settings, such as public meetings across the country called 'Save the Children', which bring the QAnon conspiracy here and place it into a local context for the audience –framed around abuse in state care, and specifically targeting Oranga Tamariki, Police, and other agencies as perpetrators of the harms that the conspiracy focuses on.

QAnon social media content and frames, shared into and on NZ-based accounts, groups, and channels is consistently the *most violent content* we discover in our research.

s9(2)(c)

---

[11] https://www.theguardian.com/us-news/2020/aug/25/qanon-conspiracy-theory-explained-trump-what-is
[12] https://www.bbc.com/news/53498434

National impacts

TDP is observing the effects of an *information void* as Covid-19 and its impacts are less focused on by institutions, including government agencies. In the absence of narrative frames, an information void develops, one that mis- and disinformation producers are now filling with alternative narrative frames which rely on racism, misogyny, queerphobia, and pseudo-law. In an environment with less counter-messaging, these are growing at pace.

Racism

Disinformation producers rely on frames that articulate, amplify, and normalise racism – and establish the right to offend and be offensive as a *pillar of free speech*. Hate speech and harmful speech are framed as free speech – allowing the generation of logic that any criticism is an attempt to 'cancel' or censor them.

This is further emphasised by requests from disinformation producers to 'chat', 'discuss ideas', or 'debate' ideas with public figures. When their requests are denied or ignored, it can be framed as a refusal to engage in 'free debate' or an attempt at cancelling.

One example of the anti-Māori racism that saturates mis- and disinformation ecosystems is the response to Matariki. 24 June 2022, Aotearoa New Zealand's first public holiday to acknowledge Matariki, the Māori New Year, generated criticism and anti-Māori racism within the ecologies studied by TDP. The public holiday saw common and widespread themes re-articulated, including He Puapua, Three Waters, Foreign Minister Nanaia Mahuta, Te Tiriti o Waitangi, co-governance, and denial of Māori indigeneity. These themes target Māori (both collectively and individually, such as Foreign Minister Nanaia Matuha) with harassment and violence. The widespread rejection of the premise of the new holiday has impacts for social cohesion.

Misogyny

Themes and frames of toxic masculinity and misogyny are commonly and normatively expressed mis- and disinformation ecologies studied by TDP. Critically, these frames are used to recruit further subscribers and to target women's participation in public life. Effects of these widespread frames are already being felt by women and gender minorities, and resulting in norm-shifting, where the use of misogyny and language of threats and violence is accepted in Aotearoa New Zealand. Misogynistic framing – particularly around the roles of men and women in families and public life – is the most common unifying frame we observe in the ecosystems studied. Examples during the period of analysis focus on abortion rights, motherhood, and the role of men, and continue to provide both tools for recruitment and for targeted harassment of women and gender minorities.

Queerphobia[13]

The publicity around Bethlehem College in Tauranga[14] and burning of Rainbow Youth[15] continue to provide narrative frames about LGBTQ+ people in Aotearoa New Zealand.

In the last fortnight, an Auckland-based misinformation producer whose main platform is Facebook attacked the visibility of LGBTQ+ people in public life, saying:

> "I am so glad that New Zealand has not stooped as low as the United States when it comes to children and sexual ideology. This seems crazy to me that people are okay with children being around hyper sexualised environments. I would love to have some LGBTQ members on the podcast with me to talk about the difference between celebrating ourselves and where the lines should be when it comes to children. If you would like to join me and chat then let's sit down for an open and honest conversation."

Another Auckland-based producer, mainly present on Telegram, was motivated by Spark's new campaign aiming to create inclusion for non-binary people[16] has attacked non-binary people. He claims Spark's campaign is "using [minorities'] stories to create falsehoods and distort social realities" which will damage the "great nation". His original post is shared into over 25 Telegram channels.

A Christchurch-based disinformation producer active on Facebook and YouTube strategically and intentionally misappropriates a Biblical passage (Pride goes before destruction, Proverbs 16:18), thus targeting Pride and LGBTQ+ communities as 'sin'. The PM's attendance at a Pride festival from years ago is highlighted, drawing her into the 'sin of Pride'.

In June, 'online' harms and discrimination against the Queer community[17] became a real display of violence with the burning of Rainbow Youth offices in Tauranga.[18] Safety, and feelings of safety, for the LGBTQ+ community are placed at risk every time dangerous, hateful, and harmful rhetoric is posited against them.[19]

Anti-establishment / sov-cit

Since 2021, TDP has analyzed the domestic growth of Sovereign Citizen (Sov-Cit) ideologies. The Sov-Cit movement emerged in the United States in the mid-1970s. Adherents view governments as illegitimate and corrupt and view themselves and living outside of the required confines of the law.[20] Since the Covid-19 pandemic, Sov-Cits in the US have pivoted into sharing Covid-19 related mis- and disinformation – including attending anti-vaccination and anti-mask events.[21] One tool of Sov-Cit rhetoric is 'paper terrorism', meaning when they get frustrated with the authorities or public institutions they retaliate with bogus legal claims that waste time and resources.[22] Sov-Cits have also been known to use violence and threats of harm. TDP is observing Sov-Cit rhetoric in Aotearoa New Zealand's mis- and disinformation ecologies – which will have negative effects on public safety, including that of frontline Police officers.

---

[13] TDP is using 'queerphobia' as an umbrella term to describe harms against members of the LGBTQ+ community.
[14] https://www.nzherald.co.nz/bay-of-plenty-times/news/taurangas-bethlehem-college-criticised-for-discriminatory-marriage-belief/ACKCSXMNTDGQ5CRCLF7AMTWZXY/
[15] https://www.rnz.co.nz/news/national/469221/rainbow-youth-tauranga-drop-in-centre-destroyed-in-suspicious-fire
[16] https://www.spark.co.nz/online/beyondbinarycode/about/
[17] https://www.nzherald.co.nz/bay-of-plenty-times/news/taurangas-bethlehem-college-criticised-for-discriminatory-marriage-belief/ACKCSXMNTDGQ5CRCLF7AMTWZXY/
[18] https://www.rnz.co.nz/news/national/469221/rainbow-youth-tauranga-drop-in-centre-destroyed-in-suspicious-fire
[19] For more, read TDP researcher Kayli Taylor's short piece on hate speech: https://thedisinfoproject.org/2022/06/18/hate-speech-in-aotearoa-new-zealand-reflecting-and-resisting/
[20] https://www.bbc.com/news/world-us-canada-53654318
[21] https://www.splcenter.org/fighting-hate/extremist-files/ideology/sovereign-citizens-movement
[22] https://www.splcenter.org/fighting-hate/extremist-files/ideology/sovereign-citizens-movement

In the last fortnight TDP has observed repeated posting on Telegram by an individual regarding their interactions with Police, and other content espoused by this person. In one visit, the individual records an interaction with Police, who advise him to take down two videos which feature graphic violence that were posted to his Telegram channel in April. The video shows two entirely separate worldviews: one connected to domestic laws and policing, and one that is inextricably entwined with and based on Sov-Cit vocabularies and beliefs.

A faux 'sheriff' van, inspired by the local 'sheriff' movement (which itself is inspired by Sov-Cit rhetoric) is celebrated on Telegram. The van says 'Stop 3 Waters', and has pictures of Chris Hipkins, James Shaw, Nanaia Mahuta, Trevor Mallard, Ashley Bloomfield, Chris Luxon, Andrew Little, Jacinda Ardern, and Grant Robertson.

TDP also analysed a letter sent to a Judge of the Supreme Court[23], and featured in the Nuremburg NZ Telegram channel, featuring a range of pseudo-science links and urging them to support their efforts to bring people in NZ who have been involved with the Covid-19 response to justice. This is the perfect example of paper terrorism. Not only is the letter, in its harassing nature and with its ridiculous demands a form of paper terrorism; but it makes reference to other behaviours that could be interpreted as the same: repeated emails to government and public health officials.

Sov-Cit rhetoric and its dismissal of Police jurisdiction could have serious effects on social cohesion, and the safety of individuals across police, government, elections, and public health. The rise in 'paper terrorism', and bombardment of law and other agencies with pseudo-legal claims will have impacts on the way these agencies operate and function.

Christchurch Terror Attack
Content warning: Discussions of the Christchurch Terror Attack

The harmful and disturbing lie that the Christchurch Terror Attack was a 'False Flag'[24] is present within the mis- and disinformation ecologies studied by TDP. *The Three Faced Killer*, a 'documentary' in three parts by Michael O'Bernicia trivialises the attacks and includes video footage from the attack – which is classified in Aotearoa New Zealand as objectionable material.[25] The second part has been released and features extended cuts from the Christchurch killer's livestream - just like Part One, which was also deemed objectionable by the Classification Office.[26] Like Part One, Part Two of the

---

[23] The letter is addressed to Chief Justice William Young, who left the Supreme Court in April 2022. The email address however, is for Chief Justice Helen Winkelmann, who is now the Chief Justice. For the purposes of this analysis, we shall label as 'a justice of the Supreme Court'.
[24] https://www.poynter.org/fact-checking/2022/what-is-a-false-flag/
[25] https://www.classificationoffice.govt.nz/news/news-items/christchurch-mosque-attack-livestream-classification-decision/
[26] Chief Censor Bans The Three Faced Terrorist, a 'documentary' about the March 15 Mosque attacks, https://www.classificationoffice.govt.nz/news/news-items/chief-censor-bans-the-three-faced-terrorist-a-documentary-about-the-march-15-mosque-attacks/

'documentary' has been banned.[27] While Part Two has been banned, like Part One and the original video footage of the attack s9(2)(c)

While TDP and Te Mana Whakaatu – Classification Office brace for Part Three, we note that subscribers to mis- and disinformation ecologies who posted the video will likely have observed it – deeply disturbing material that displays, at least in part, the attack of 15 March. The impacts on mental health, on perceptions of violence, and desensitisation[28] are myriad.

The lie that the terrorist attack was a false flag is harmful to the communities and whānau most affected by the violence, further generating harm against the Muslim community for an event that has already caused significant harm. The continued accusation that the terrorist attack was orchestrated by the Government serves only to cement distrust of the state and institutions. This will have long-tail effects on the way subscribers to mis- and disinformation ecologies interact with public institutions.

---

[27] https://www.classificationoffice.govt.nz/news/news-items/acting-chief-censor-bans-video-featuring-the-march-15-mosque-attacks/

[28] https://www.apa.org/topics/video-games/violence-harmful-effects

Appendix One: Definitions

Misinformation: "false information that people didn't create with the intent to hurt others"
Disinformation: "false information created with the intention of harming a person, group, or organization, or even a company"
Malinformation: "true information used with ill intent"[29]

Conspiracy theory: purported explanations which cite a conspiracy at the salient cause of some event or phenomenon.[30]

Dangerous speech: "dangerous speech is any form of expression (e.g., speech, text, or images) that can increase the chances that its audience will condone or participate in violence against members of another group."[31]

Hallmarks of dangerous speech:
• Dehumanisation
• Coded language
• Accusation in a mirror
• Threat to group integrity or purity
• Assertion of attack against women and girls
• Questioning in-group loyalty

[29] Berentson-Shaw J and Elliot M. *Misinformation and Covid-19: a briefing for media*. Wellington: The Workshop; (2020).
[30] Dentith MRX. Conspiracy theories and philosophy: bringing the epistemology of a freighted term into the social sciences. In JE Uscinki (ed.) *Conspiracy Theories and the People Who Believe Them*. Oxford: Oxford University Press; (2018).
[31] The Dangerous Speech Project, *Dangerous Speech: A Practical Guide*: 19 April 2021 https://dangerousspeech.org/guide/

Appendix Two: TDP's work to report and flag content to minimise harm to New Zealanders

9(2)(ba)

Reporting of online harms and threats against the Covid-19 Response: Fortnightly Update

SENSITIVE
25 July 2022
For the period 8 July to 22 July 2022

Kate Hannah, Sanjana Hattotuwa, Kayli Taylor
The Disinformation Project (TDP)

For further inquiries or should this report fall within the scope of an Official information Act request, contact the Disinformation project Director, Kate Hannah

s9(2)(a)

Executive Summary

- Aotearoa New Zealand's disinformation ecologies are a complex and shifting phenomena that is having and will continue to have impacts on human and national security.
- Social media groups and channels studied have subscriber numbers in the hundreds of thousands: 380,000 subscribers across 161 channels on Telegram; 956,685 followers across 95 Facebook pages; 220,442 followers over 47 Instagram accounts.
- Disinformation producers continue to target vaccinations as harmful, masks as ineffective, and all public health measures as ridiculous.
- Monkeypox has been imbricated into conspiratorial thinking and denialism.
- The image of the PM, other senior government officials, and youth MPs unmasked is a gift to disinformation ecologies and is an accelerant for worsening information disorders in Aotearoa New Zealand.
- Identity-based harassment, including racism and misogyny, continue to rise in the ecologies studied by TDP.
- TDP witnessed more Sov-Cit rhetoric this fortnight, including threats to bring the PM to trial.
- TDP is increasingly concerned about the threat of stochastic terrorism in NZ.
- The white supremacist 'Great Replacement Theory' is referenced without critique in domestic Telegram channels by NZ-based producers.
- TDP notes how easy it is to move from domestic anti-vaccination and anti-mandate Telegram channels to channels promoting violent extremism and other harms.
- Sri Lanka's unrest has captured the attention of mis- and disinformation ecologies studied by TDP. The protest activity is held up as an example of how to revolt against government – including in Aotearoa New Zealand.
- Former Japanese President Abe's assassination was reported on Telegram faster than many mainstream media outlets picked it up, and later framed conspiratorially.
- Mis- and disinformation producers oppose Three Waters reform, muddy the issue, and protest visibly.
- Russian disinformation continues to be shared amongst ecologies studied by TDP, including domestically produced content in Russian.
- Mainstream media has promoted misinformation, and thus bolstered its producers.
- Schools are, as we have described over the past year, increasingly contested sites where efforts to increase social cohesion such as the new history curriculum are poised to become embattled.

Health disinformation

TDP observes in studied ecologies health disinformation related to Covid-19 denialism and minimisation, anti-public health rhetoric, and anti-vaccination messaging. The emergence of Monkeypox has also drawn the attention of mis- and disinformation ecologies.

A large and popular disinformation group which focuses on women and families' features another disinformation narrator claiming "You can statistically show that the vaccines have been increasing the deaths" and represents the Covid-19 vaccine as a "device that changes the way our immune system works." These represent explicit claims that what is being named as Covid-19 related deaths are instead vaccine-related. Another high profile and popular woman disinformation producer posts videos alleging that Covid-19 vaccines cause myocarditis, heart issues, heart attacks and essentially, kill people. This denial of the effects of Covid-19 is highly palatable as it offers a neat explanation for excess death.

s9(2)(g)(i)

Anti-mask discourses and the promotion of the misuse of mask exemptions are widespread. Multiple disinformation producers appear to be building up to a crescendo that will be unleashed in its full force if/when stronger mask mandates are announced by Government. Any capitulation is also packaged as evidence of their power and influence — ie the strong advice to schools that falls short of an actual mask mandate is understood within the location of study as evidence of the fear of their growing movement.

Media reporting on Monkeypox cases in Aotearoa New Zealand[1] has been recognised within the ecosystems studied by TDP. Monkeypox has been immediately drawn into anti-public health measures across mis- and disinformation ecologies studied by TDP. A poll on Telegram from an alternative 'news' organisation which produces and promotes disinformation highlights resistance to a lockdown, as well as the belief that the New Zealand government will introduce a lockdown as a system for control (of both people, and virus). Broadly, Telegram's reception to Monkeypox is discourse is exclusively ridicule and rejection. The thrust of comment responses to this poll includes anti-vaccination, anti-mandate, anti-government, Covid-19 denialism, and Monkeypox denialism. Monkeypox has been drawn into the same operation of conspiratorial thought in which Covid-19 is viewed.

This graph compares interactions of the Unite Against Covid-19 Facebook page against that of the high-profile woman disinformation producer whose livestream content was the most popular during the Parliamentary occupation.

---

[1] New Zealand's first case of monkeypox detected in Auckland, https://www.stuff.co.nz/national/health/300633595/new-zealands-first-case-of-monkeypox-detected-in-auckland

As the graph shows, since January 2022 the misinformation producer has received nearly *four times* the number of engagements than the Unite Against Covid-19 page. While the majority of this was over February-March 2022, this individual remains slightly higher interactions in May and June 2022.

A maskless Prime Minister
Against a backdrop of worsening information disorders, rising Covid-19 infections and re-infections, and a health system on the verge of collapse,[2] an image of a mask-less Prime Minister, Governor General, other MPS, and youth MPs is a *gift* to mis- disinformation ecosystems and an accelerant for worsening information disorders in Aotearoa New Zealand.

TDP has written over 30 pages, summarising the response from both mis- and disinformation ecologies and those not subscribed, including politicians – current and former – and academics. In the interest of brevity, we will summarise a few points below.

Mis- and disinformation ecologies latched on immediately, and significantly, with a variety of responses. Some labelled the Prime Minister hypocritical, others used it as an opportunity to allege that masking is not effective, some highlighted other posts from Youth MPs in which individuals are seen mask-less, and some labelled the PM's rhetoric as "Do as I say, not do as I do". No mis- and disinformation producer has achieved the level of undermining of public health measures as this image has. Within mis- and disinformation ecologies, it has further undermined the government's own public health guidelines, policies, and communications in ways that TDP expects to have longtail effects.

Discourse on Twitter shows 221 tweets on the subject, reaching a potential 46,920 followers. These tweets generated 2662 retweets (including quote tweets). TDP noted last week that trust in government and public health organisations from those *with previously high trust* in these groups is being eroded. This image further erodes this trust – with implications on how future public health responses are likely to be received.

Disinformation ecologies already had anti-public health measure views, which manifested into anti-masking rhetoric. TDP stresses that all future expressions of anti-mask sentiment will be appreciated by the PM's maskless photo on social media. Thus, pushback against mask use is *strengthened* by the PM's own actions.

---

[2] Covid-19 NZ: Why the rising tide of cases doesn't tell the whole story, https://www.stuff.co.nz/national/explained/129256918/covid19-nz-why-the-rising-tide-of-cases-doesnt-tell-the-whole-story

In sum, the image presents a serious and unprecedented issue in domestic information disorders – the magnitude of which is yet to fully be seen. This image, alongside decreasing public and official communication creates a new foundation for mis- and disinformation to thrive.

Beyond health disinformation – other trends of the ecosystem

Identity-based targeting and harassment

TDP has repeatedly pointed to the ways in which Covid-19 mis- and disinformation ecosystems are drawing people towards conservative ideologies, far-right views, and racism.

This fortnight, disinformation producers across Telegram and Facebook promoted an online petition against calling this country Aotearoa, which was hosted on the anti-Māori Hobsons Pledge website. Racism is deeply intertwined with the disinformation ecologies studied by TDP – with impacts on all of Aotearoa New Zealand and its efforts for social cohesion. There has been a notable increase in both antisemitic and Islamophobic content in commentary in the period of study.

Additionally, themes and frames of toxic masculinity and the operations of misogyny are present and highly volatile across the mis- and disinformation ecologies studied by TDP. For example, the Freedom and Rights Coalition promotes "real men" joining a "Million-Man March" in Auckland, Wellington, and Christchurch with language such as "Men of this nation will gather up their wives, sons and daughters and say, "Let's make history and stand for our freedoms, our rights and let's get our nation back." The language is heteronormative, erases gender diverse identities, and frames a highly misogynistic expectation that men need to "fight" to save the nation from its current "collapse". Disinformation producers who are women are also involved in the production and promotion of material with harms to gender equality and the lives and safety of women and gender minorities.

Sovereign Citizen

Since 2021, TDP has borne witness to the domestic growth of Sovereign Citizen (Sov-Cit). The Sov-Cit movement emerged in the United States in the mid-1970s. Adherents view governments as illegitimate and corrupt and view themselves and living outside of the required confines of the law.[3] Since the Covid-19 pandemic, Sov-Cits in the US have pivoted into sharing Covid-19 related mis- and disinformation – including attending anti-vaccination and anti-mask events.[4] One tool of Sov-Cit rhetoric is 'paper terrorism', meaning when they get frustrated with the authorities or public institutions they retaliate with bogus legal claims that waste time and resources.[5] Sov-Cits have also been known to use violence and threats of harm. TDP is observing Sov-Cit rhetoric in Aotearoa New Zealand's mis- and disinformation ecologies – which will have negative effects on public safety, including that of frontline Police officers. TDP notes that Sov-Cit rhetoric is now strong enough offline to make mainstream media news.[6]

s9(2)(a)

Other disinformation producers continue to travel the country promoting Sov-Cit ideas and spreading conspiratorial thought.

---

[3] https://www.bbc.com/news/world-us-canada-53654318
[4] https://www.splcenter.org/fighting-hate/extremist-files/ideology/sovereign-citizens-movement
[5] https://www.splcenter.org/fighting-hate/extremist-files/ideology/sovereign-citizens-movement
[6] https://www.stuff.co.nz/national/crime/129313220/trials-and-tribulations-during-mans-troubled-court-appearance

A high profile former mainstream journalist disinformation producer interviews another a fringe disinformation producer who has been promoting Sov-Cit rhetoric and his encounters with Police on his Telegram channel. This narrator and second speaker 'interview' technique provides and powerful form of dangerous speech that validates fringe ideas through the medium of a 'news' style interview conducted by someone with status within that space.

In the post accompanying the video, the interviewer says "we must create change at a local level… Take out your cameras, and make a lot of noise! Demand Change!" The video itself amplifies Russian disinformation, vaccine related disinformation, and re-features previous content from the interviewer alleging children are dying from the Covid-19 vaccine. Towards the end of the video, the interviewee questions the Police, vaccine clinics, and anyone associated with power. When asked to say something to PM Ardern, the interviewee says, "Resign and get ready for trial."

Methods of engagement and protest
TDP observes a variety of methods of engagement and protest within the disinformation ecologies we study. Some of these are tools currently being implemented, like public protest and use of the mainstream media. Some are hinted at – realities that TDP is increasingly concerned we will witness in Aotearoa New Zealand. One example of this is stochastic terrorism.

On 8 July, following the release of a publication called 'The Hard Reset', the Counterterrorism Group (CTG) in the United States released a flash alert warning that the publication would "almost certainly" increase extremist violent attacks across the country.[7] They noted that the motive behind the publication was to encourage white supremacists and anti-government individuals to take violent action.

s9(2)(c) TDP has read the document and agrees with CTG regarding the tone and thrust of the content. We cannot make an assessment about potential offline consequences of the availability of the document in a domestic context. However, given the content of the document and the irrigated path dependencies established by prior TVEC content, s9(2)(c)

We are observing the very real threat of stochastic violence and terrorism. The ease of access to 'The Hard Reset' and other materials across s9(2)(c) highlight that Aotearoa New Zealand is reaching a point where kinetic harms including, but not limited to, stochastic terrorism are *inevitable*.

The Great Replacement Theory in Aotearoa New Zealand
One disinformation producer, s9(2)(c) signals The Great Replacement Theory's central thesis: that the white population, suffering from declining birth rates, is being 'replaced'

---

[7] https://www.counterterrorismgroup.com/post/flash-alert-high-risk-of-violence-with-the-publication-of-the-hard-reset-a-terrorgram-publication

by immigrants. He names it in the context of the Dutch farmers protests, alleging he heard about dropping birth rates amongst white people in the Netherlands, and high immigration flows. The Great Replacement Theory (GRT) inspired the Christchurch terrorist and the Buffalo mass-shooter. This disinformation producer's rhetoric and framing is the same ideological framework, and vocabulary, as the aforementioned terrorists.

Domestic Telegram ecologies are one step from violent extremist promotion
TDP's study focuses on s9(2)(c) who cluster around anti-vaccination and anti-mandate messaging. These channels regularly link to, share, or direct people towards terrorist and violent extremist content, violent extremist, child sexual abuse material, and The Great Replacement Theory repositories. There is no guard, friction, oversight, or control over the production and propagation of this material – and no geographical containment. What is produced in the United States is instantly discoverable by those in Aotearoa New Zealand who are imbricated within mis- and disinformation ecologies.

Sri Lanka
The socio-political developments in Sri Lanka have caught the attention of disinformation ecologies studied by TDP. The events are held up as an example of how people could revolt against the government in Aotearoa New Zealand, how food shortages in Sri Lanka preface the same here, and how civil unrest in Sri Lanka will also be experienced in Aotearoa New Zealand. Sri Lanka's conundrums and issues are instrumentalised without context or historic framing to suggest civil and political unrest could be achieved in Aotearoa New Zealand in the same way.

President Abe's assassination, represented on Telegram
Former Japanese President Abe's deadly shooting was captured quickly by ecologies studied by TDP. The first post on Telegram was posted several minutes prior to any wire report seen by TDP on Twitter, or before reporting from the *New York Times*. Consequently, TDP views Telegram as a real time news network, reporting entirely independently from wire news reporting and mainstream media. Later framing of the assassination is dominated by tropes that he was close to Putin and opposed vaccine mandates, the World Economic Forum, globalists, and the World Health Organisation, and that his assassination is therefore the work of pro-vaccine conspirators.

Opposition to Three Waters
Multiple content producers within disinformation ecologies studied by TDP have strong opposition to Three Waters reform. This includes organising physical protest outside the Local Government New Zealand (LGNZ) meeting in Palmerston North – which was shared across multiple clusters within our ecologies. Three Waters reform is almost exclusively framed as concerns of Māori control or ownership over water – a sign of the racism present within ecologies, as well as the way the *debate* about Three Waters has overshadowed the actual issue.[8]

---

[8] https://thespinoff.co.nz/live-updates/21-07-2022/commentary-around-three-waters-has-overshadowed-need-for-change-ardern
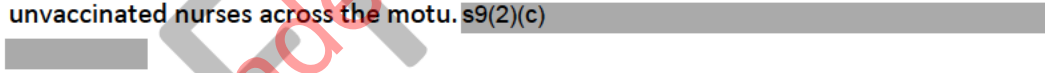
The spread of Russian disinformation

TDP has noted the spread of Russian disinformation and propaganda throughout Aotearoa New Zealand's information ecosystems, particularly since the invasion of Ukraine in February 2022. In a paper released in May, we described how, by the end of March 2022, every domestic Telegram channel studied by TDP had pivoted to a nearly exclusive framing of the Ukraine war through pro-Putin and pro-Kremlin frames. This reflects trends observed by Microsoft, who reported in June that after December 2021, Russian propaganda consumption in New Zealand increased by over 30% relative to consumption in Australia and New Zealand. Critically, pro-Kremlin and pro-Putin content in the Russian language is being produced in New Zealand, via open channels and closed groups such as https://www.facebook.com/groups/VladimirPutinFanClubNZ/discussion/preview. This includes content which seeks to fundraise for Russian causes, potentially in violation of sanctions and terrorism laws.

s9(2)(c)

Mainstream media promoting misinformation

This fortnight, the mainstream media promoted a puff piece for a group of unvaccinated nurses begging to return to work amidst the collapsing health system.[9] The piece failed to recognise the cluster was organised by a prominent disinformation group, which has conspiratorial ideas about the New World Order, believes the vaccine will cull the population, and advocates for Nuremberg trials for doctors, academics, and politicians. The group of nurses and its parent network, following the success of the piece have scaled up rapidly, organising action plans for unvaccinated nurses across the motu. s9(2)(c)

s6(d)

---

[9] https://www.stuff.co.nz/southland-times/news/129197272/plea-by-unvaccinated-nurses-to-return-to-work

s6(d)

Appendix One: The Disinformation Project and our field of study

Disinformation is a threat to democracy,[10] stable governance,[11] and human life.[12] Since the start of the Covid-19 pandemic and its associated infodemic,[13] disinformation and its impact on people and society in Aotearoa New Zealand has grown. The Disinformation Project (TDP) has been analysing this since February 2020, paying particular attention to the volume, velocity, and vector of information. We use daily data collection and analysis (which form the basis of this summary report), use computational and manual tools to scan open-source social media post and commentary across a wide range of social media platforms, websites, and media/alternative media organisations. The information landscape studied is developed using 'snowball' techniques, which means that we have expanded the inclusion of pages, groups, and channels only when they are signalled by existing locations of study.

Within the social media ecologies studied, key individuals and groups producing mis- and disinformation capitalise on growing uncertainty and anxiety amongst communities, related to Covid-19 public health interventions, including vaccination and lockdowns, to build fear, disenfranchisement, and division. Mis- and disinformation is also particularly targeting and scapegoating already marginalised or vulnerable communities – for whom distrust of the state is the result of intergenerational trauma and lived experience of discrimination or harm, which can increase engagement with conspiratorial explanations and disinformation. Over the past two and a half years of research, TDP has developed a thorough and balanced understanding of the harms that mis- and disinformation and 'dangerous speech' present to social cohesion, freedom of expression, inclusion, and safety. (See Appendix One for our definitions of these terms).

The landscape studied – Covid-19

The landscape studied originated as locations engaged in mis- and disinformation related to Covid-19 and the Covid-19 response, and our study, the type of content produced and shared within this landscape has shifted over time, so other narratives and themes within this landscape now form part of our analysis. For the purposes of this reporting, we focus on online harms and threats against the Covid-19 response, including people and places associated with the Covid-19 response. This includes covid denialism, covid minimisation, anti-vaccination messaging (which is increasingly spilling from vaccination against Covid-19 into other types of vaccination, including vaccination for tamariki), and anti-mandate/anti-public health measures messaging. These are the four dominant types of messaging that TDP observes in its study of the mis- and disinformation ecologies within Aotearoa New Zealand.

---

[10] https://www.europarl.europa.eu/RegData/etudes/STUD/2021/653635/EXPO_STU(2021)653635_EN.pdf
[11] Ibid.
[12] https://www.scientificamerican.com/article/covid-misinformation-is-killing-people1/;
https://www.axios.com/2022/04/21/barack-obama-disinformation-social-media
[13] https://www.who.int/health-topics/infodemic

Appendix Two: Definitions

Misinformation: "false information that people didn't create with the intent to hurt others"
Disinformation: "false information created with the intention of harming a person, group, or organization, or even a company"
Malinformation: "true information used with ill intent"[14]

Conspiracy theory: purported explanations which cite a conspiracy at the salient cause of some event or phenomenon.[15]

Dangerous speech: "dangerous speech is any form of expression (e.g., speech, text, or images) that can increase the chances that its audience will condone or participate in violence against members of another group."[16]

Hallmarks of dangerous speech:[17]
• Dehumanisation
• Coded language
• Accusation in a mirror
• Threat to group integrity or purity
• Assertion of attack against women and girls
• Questioning in-group loyalty

---

[14] Berentson-Shaw J and Elliot M. *Misinformation and Covid-19: a briefing for media*. Wellington: The Workshop; (2020).
[15] Dentith MRX. Conspiracy theories and philosophy: bringing the epistemology of a freighted term into the social sciences. In JE Uscinki (ed.) *Conspiracy Theories and the People Who Believe Them*. Oxford: Oxford University Press; (2018).
[16] The Dangerous Speech Project, *Dangerous Speech: A Practical Guide*: 19 April 2021 https://dangerousspeech.org/guide/
[17] Ibid.

Appendix Three: TDP's work to report and flag content to minimise harm to New Zealanders

9(2)(ba)

Appropriate frameworks for social media analysis for New Zealand

June 2022

BRAINBOX

www.brainbox.institute

Commissioned by the Department of the Prime Minister and Cabinet.
This report is intended to suggest pragmatic steps toward effective and appropriate social media analysis for public policy purposes within New Zealand.

Context

Recent years have seen growing demand from some members of the New Zealand public and media for increased government capture and analysis of internet-based communications occurring via social media websites and apps. These demands began in earnest in the wake of the 15 March 2019 Christchurch Terror Attacks and were reinvigorated throughout the Covid-19 pandemic period, peaking in proximity to the prolonged occupation of the lawn outside Parliament.

Participants at this protest were substantially inspired and organised by communications occurring on social media. Some of these communications almost certainly contained deliberately false information, strategically propagated to inspire beliefs and actions in accordance with the propagandists' objectives. Such communications are often referred to as "disinformation". When further propagated by well-meaning audiences, they are referred to as "misinformation".

There is a strong case to be made for the establishment and support of a diverse, multidisciplinary civil society-led institution to conduct ongoing analysis of social media-based communications (henceforth referred to as 'social media analysis' or 'SMA') for the purpose of monitoring and analysing potential disinformation and misinformation. Overseas, similar institutions have contributed materially to enhanced security.

Brainbox also participated in-person at the Atlantic Council Digital Forensic Research Lab's 360 Open Summit in June 2022 – the premier global forum for disinformation expertise. During this time, we tested and corroborated many of the positions set out in this report.

CONTENTS

Executive Summary

The case for systematic analysis and scrutiny of the communications exchanged on social media (henceforth referred to as social media analysis or 'SMA') is straightforward. In the best-case scenario, effective surveillance of social media communications can produce useful insights about the extent to which disinformation is occurring. Equipped with these insights, different stakeholders across society can take effective action to reduce the harms that disinformation may cause.

How this is carried out

The fundamentals of SMA are simple. Data – primarily public communications made by users on social media platforms – is collected and then analysed. These communications are often text-based, and subsequently can be analysed computationally in conventional ways, like keyword searching and the counting of visible metrics of engagement by audiences.

Although facilitated by statistical programs and techniques, this kind of activity still requires extensive manual analysis and the exercise of significant human judgement. While there is a growing body of research that aims to develop automated techniques to detect or analyse disinformation without need for manual human intervention, such technologies remain unreliable.

Detection and analysis are hard

There are numerous difficulties inherent in detecting and analysing disinformation in social media-based communications. The two most significant are:

- Determining whether communications meet the criteria of disinformation. Despite a consensus around a theoretical definition, many of the boundary criteria for that definition are difficult to objectively assess externally, such as the intention of the communicator. Others may be uncertain, such as the accuracy of the information. As a result, determining whether a communication is disinformation can be highly subjective – neutral observers might disagree in good faith. Extrapolated over enormous datasets, this undermines the reliability of findings.

- Accessing high quality data. Access to data is fraught, and even when large datasets can be prepared, this data may be of poor quality for the purposes of the analysis. As a result, researchers are compelled to extrapolate from small datasets or study platforms and issues which facilitate easy access to data, whether or not they are the most pressing subjects. Without prudence and integrity, this materially undermines the reliability of the findings. This difficulty is exacerbated by the growing number of communications across multiple online apps and websites.

Government faces unique challenges

There is no doubt that parts of the New Zealand government, such as Police, already conduct SMA to some extent. In some cases, this is a necessary and useful part of the government carrying out its duties, including to safeguard the rights of citizens. Outside law enforcement, a range of government agencies also use off-the-shelf SMA products to monitor engagement with State communications on social media.

Nevertheless, the government has obligations to act legally and properly. These obligations create practical barriers for government officials who wish to carry out effective capture and analysis of internet-based communications. For example:

- The Terms of Service for most social media platforms prohibit the large-scale 'scraping' of data without their express oversight and permission. Some jurisdictions have ruled that there are implicit exceptions for non-commercial research, but the practice largely remains in a legal grey area. While more broadly accepted in academia, such techniques are more controversial if undertaken by government researchers or contractors.

- Although it is true that disinformation occurs in publicly accessible social media-based communications, many of the most impactful forums for disinformation are not publicly accessible, e.g. closed Discord channels, WhatsApp groups, Telegram channels, or private Facebook groups. Gaining access to these requires that a researcher behave deceptively. This bears resemblance to orthodox espionage tradecraft. The regulation of New Zealand government agencies and public servants strictly controls the circumstances and manner in which government officials may conduct this kind of activity.

But even if government SMA is scrupulously legal and ethical, it will unavoidably attract negative public attention due to a host of anxieties around privacy, free expression, and government influence over public discourse. While the extent of this backlash can be mitigated to some degree by keeping SMA limited in scope and fully transparent, it is fundamentally unpredictable and runs the risk of undermining trust in government, further radicalising at-risk users, driving away potentially useful partner organisations, and delegitimising future efforts to combat disinformation.

Towards a hybrid governance model

Acting on its own, there is no viable way for the New Zealand government to access the benefits of SMA for disinformation monitoring and mitigation. Rather, a non-government entity with appropriate governance structures and funding security is the best vehicle for this. This entity can formally or informally incorporate a multistakeholder arrangement, with stakeholders potentially including civil society, academia, industry, and government itself. In addition to acting as a mechanism for balancing the many important rights and concerns inherent in this undertaking, a civil society group is likely to be a more appealing partner for social media platforms and other entities that are reluctant to directly collaborate with state governments.

Non-government organisations in other jurisdictions have produced world-leading research, conducted crucial outreach efforts, and provided valuable insights and advice to lawmakers. And New Zealand has a number of unique features – among them mātauranga Māori, Te Tiriti obligations, a comparatively high level of social cohesion and media trust, its geopolitical location in the Indo-Pacific, and a highly specific socioeconomic milieu – that provide compelling reasons for undertaking New Zealand-based work rather than importing experts and conclusions from other jurisdictions.

Drawing on a report on a similar topic by the Institute for Strategic Dialogue (ISD), Brainbox proposes the following design principles to ensure that this civil society group is able to maintain public credibility, analytical rigour, and policy relevance:

- **Full integration into civil society**, bringing together a wide range of participants.
- **Data access as a priority**, taking advantage of every source and platform.
- **Cross-platform focus**, studying a range of platforms and the interactions between them.
- **Continual self-assessment and development**, improving its capabilities and tools in response to new requirements and research.
- **Explicability at all stages**, making sure that both final outputs and analytical processes are accessible and understandable.
- **Insight from all sources**, bringing together useful frameworks and information from SMA, sociology, psychology, and community representatives.
- **Te Ao Māori centrality**, ensuring Te Tiriti obligations are met and mātauranga Māori is respected.

New Zealand has the opportunity to learn from overseas successes. It can model best practices in addressing these important and highly charged issues in a way that is responsible, rigorous, and fully engaged with academia, civil society, Māori perspectives, and the broader public.

Part I: Analysing social media disinformation

Why study manipulative communications?

Although the role of the internet and technological architectures has recently reinvigorated public and political interest in such matters, social media is only the latest theatre for propaganda. Disinformation is a subset of propaganda, which has been studied deliberately since at least the 1920s. Research from this period gave rise to the field of behavioural psychology that forms the foundation of modern marketing. Since then, propaganda (and thus disinformation) has been a persistent feature in statecraft, warfare, politics, and business – even if public interest in the subject has ebbed and flowed.

Nevertheless, the internet and social media have revolutionised traditional influence practices. They have enabled asymmetrical propaganda activities to be carried out on a global scale at much less cost, with little regard for time, distance, or local laws. Most people now recognise that online manipulation has been leveraged to affect societies in almost every material way: politics, beliefs, values, identities, purchasing habits, and more.

Consequently, social media-based communications are now routinely and systematically analysed by researchers across academia, business, and civil society. Much of this research focuses exclusively on disinformation and its effects. Some government agencies (e.g. Police) also conduct SMA to some degree.

The fundamentals of social media analysis

The fundamentals of SMA resemble other conventional areas that make use of internet data – particularly marketing and advertising, both commercial and political. Data is first collected and then analysed.

Data Capture & Collection

- Researchers gather whatever data they can access. With enough money, licences to access large commercial datasets can be purchased from third-party data brokers, or from social media platforms themselves. For more targeted or low-budget research, data can be collected (often skirting the terms of service) directly from websites and apps. There is growing momentum toward researchers open sourcing their computer programs for conducting this kind of analysis in order to build civil society capacity and avoid duplication of resources.

- Data gathered typically consists of publicly available communications posted by social media users; both their content and what is called 'metadata': information about the communications such as time posted, number of "likes", "shares", "retweets", and "impressions".

- Some researchers may attempt to gain access to non-public spaces, such as closed Telegram channels and private Facebook groups, in order to collect data on the communications therein. This often entails some level of deceptive behaviour. There are also recent examples of the use of likely unlawful techniques by actors with pro-social intent to access and disclose information that exposes harmful behaviour, for example among white supremacist groups.

Data Analysis

- Researchers scrutinise the data to develop inferences about what it says and how it can be leveraged. Some level of manual analysis is almost always necessary, if only to verify and legitimate the outputs of automated analysis systems.

- These automated systems are typically designed to parse text and extract insights. This can be as simple as searching for key words in communications, or as complex as estimating the 'sentiment' of social media posts associated with certain topics. The more complex the analysis, the greater the risk that automated systems can mislead – either through biased construction, or failure to capture communications' full context.

- Researchers often map networks of user-accounts that are publicly communicating with each other or sharing the same content. This can identify the most prolific communicators, and to some degree the most influential accounts. It can also give some indication of whether communications are gaining traction with new audiences. This mapping exercise can be confounded if activity is occurring across multiple platforms and websites, where researcher access is limited only to particular platforms.

Part II: Limitations of the current landscape

Current research into social media disinformation is neither fully comprehensive nor conclusive, primarily due to three fundamental obstacles to effective SMA: **Definitional difficulties, barriers to accessing or processing relevant data,** and the **overwhelming complexity** of the systems and influences in question.

Definitional difficulties

Disinformation has various theoretical definitions in academic and policy contexts. It is generally regarded as being false information created or distributed intentionally, sometimes with intent to cause harm. Each of these criteria creates practical difficulties:

- **Intent**: Unless it is explicitly stated, intent must be inferred from context. Factors such as complexity, anonymity, cultural variance, and deliberate obfuscation make these inferences challenging. Accurately inferring intent is time consuming, prone to bias, and sometimes impossible.

- **Falsehood**: Most definitions agree that disinformation must be false. However, this criterion poses three challenges. Firstly, complex issues often cannot be reduced to a binary of true or false. Secondly, some claims are verifiably true (i.e., they are empirically observed facts), but presented in a skewed frame or stripped of important context. Finally, many statements that may reasonably be called disinformation are ambiguous, cloaked in irony, or simply non-falsifiable. The concept of truth is also politically contested in various ways unique to New Zealand (for example, discussions around mātauranga Māori).

These difficulties lead to several issues which undermine the quality of the analysis produced.

- First, the application of the theoretical definition around intent and truth is highly subjective, increasing the risk that researcher bias shapes research findings.

- Second, it leads to the adoption of proxies or working definitions that do not adequately match the theoretical definition as previously outlined, but rely on more easily observable indicators. This means that headline findings about "disinformation" can also be misleading.

- Finally, even if a satisfactory in-practice definition of disinformation can be developed, it is very difficult to develop automated systems that can consistently apply it – a necessity, given the vast quantities of data that must be processed.

Barriers to accessing/processing relevant data

- Limited access to data constrains the quality of research. While all science encounters this problem, the study of internet communications is particularly frustrated by the fact that there is an essentially limitless quantity of data in the hands of private companies which researchers cannot easily gain access to.

- Terms of Service for platforms, which usually constrain wholesale extraction of data, can lead researchers without a relationship with the platform to either limit sample sizes or skirt the TOS.

- This incentivises the study of data that is relatively easy to collect, like public posts on large social media platforms. By contrast, many of the most egregious and impactful examples of disinformation likely occur on smaller and less scrupulous websites, forums, image boards, private groups, or generally places where data on the communications taking place is far less accessible.

- Automated tools for analysing audio-visual content are significantly less accurate than those available for text. This can lead researchers to neglect this category of content, which is thought to be a highly significant one in the spreading of disinformation.

- It is not uncommon for researchers to withhold their methodologies or datasets in the interests of safety and security. While this may be justified in some cases, it prevents effective scrutiny of their results or methods. This makes it difficult to have high confidence in research findings, to identify and learn from mistakes, or to suggest improvements, which hampers scientific progress.

- Additionally, the pace of current inquiry is not conducive to peer review. There is genuine and justified urgency to try and produce results and recommendations in time for them to be useful – e.g. before an election, or within the timeframes of a vaccination drive. However, this all but eliminates opportunity to replicate a study within relevant timeframes, and many if not most studies on online disinformation have likely never been subjected to a single replication attempt.

Overwhelming complexity

- False claims and damaging narratives are spread between platforms by countless formal and informal networks of users – rendering each platform both its own environment and a node in a vast, ever-shifting 'information ecosystem'.

- Platforms are constantly developing. Users, moderation policies, Terms of Service, and even technological foundations can change rapidly, making it more difficult to rely on past research or methods as a guide.

- There are a huge number of vectors for false information: user posts, ads, news articles, memes, livestreams, and many more. What disinformation looks like in practice is different for each vector and platform, making it difficult to study them all with one approach.

- New platforms are constantly emerging (such as Yubo, the platform used by the recent Uvalde shooter), and their place in networks and 'information ecosystems' often takes time to become apparent. New platforms also typically have less developed transparency processes and are thus more difficult to study than their more established counterparts.

There are still numerous contemporary examples of useful open-source research on topics of online disinformation. For a selection of case studies, see **Annex A: Case Studies**.

Technological solutions are unreliable

We caution against enthusiasm towards advanced technological solutions for monitoring or moderating online disinformation. Many researchers and companies claim to have developed machine learning (ML) systems for the automated detection of 'fake news' or 'deceptive content', but in practice these systems tend to use extraordinarily blunt metrics and have unacceptably high

margins of error even when deployed in carefully controlled lab environments.

While there will undoubtedly be some role for ML going forward, even its most advanced applications have significant limitations. Social media platforms have been using ML to automate aspects of moderation for many years, with mixed results; while companies regularly release reports on the swathes of rule-breaking content removed by these systems, large volumes of misinformation on their services continue to escape detection – even on topics which have seen great focus and intensive fact-checking efforts, such as Covid vaccines.

Ultimately, there is currently no technological solution for SMA that fully mitigates the need for significant manual work by qualified personnel with adequate comprehension of cultural factors among relevant communities.

Common research techniques are not suitable for high-tempo work

In this assessment, we have been considering how governments might use SMA to inform operational decisions and guide public messaging. It is therefore important to emphasise that much of the best work described in academic or NGO literature is performed retrospectively. It requires significant resource investments and may be poorly suited to real-time decision-making. By contrast, we identified one method for high-tempo SMA that, while potentially effective, raises significant risks for legality, proportionality, and human rights protections.

A resource-efficient SMA model for situational awareness

Studies consistently show that a minority of users are responsible for most of the communications within any social media group or community of interest, and this is also true in disinformation contexts. While conspiratorial narratives are typically generated by a cyclic exchange between influencers and the wider conspiracy community, influential framings and claims will typically pass through these key actors.

As a result, identifying and monitoring key 'influencers' in anti-vaccine and other conspiratorial communities in New Zealand could theoretically provide a tolerably accurate and timely indication of the narratives being discussed in and disseminated by these communities.

An initial expenditure of cash, time, and expertise would be required to identify these key actors (in addition to those already identified by efforts to date, such as the Parliament occupation's so-called 'disinformation dozen'), but subsequent monitoring would be relatively low cost – likely requiring only a small number of personnel checking in at regular intervals to parse chatter and record the emerging themes.

However, Brainbox recommends against this approach for the following reasons:

- It is highly likely specifically monitoring individuals would amount to domestic surveillance. As such, it would need to be conducted pursuant to relevant legislative and oversight frameworks.

- Even if lawful, such activities may nevertheless be inconsistent with international human rights norms and invite widespread condemnation, undermining New Zealand's international diplomatic position. It may be that the Government wishes to make the case for conducting such monitoring by agencies without a law enforcement function, however it is critical that agencies performing such monitoring make that case directly, and do not edge into unlawful or unjustified surveillance under the guise of SMA.

- Many high impact disinformation influencers raise themes around distrust of government, allege enhanced and secret state surveillance, and persecution of people based on expression of minority viewpoints. If the government were to engage in this kind of behaviour it would undermine public trust and confidence in the government while enhancing the standing of those "disinformation influencers" by providing actual or perceived evidence for their claims.

Part III: Government-specific challenges

The New Zealand government has limited funds, personnel, and expertise to dedicate to social media analysis. Leading SMA firms charge high prices for access to their expertise and systems, which may be difficult to justify given the uncertainty of outcomes in this area.

In addition, systematic government capture and analysis of internet communications would arguably amount to unjustified government surveillance. This speaks to both of the two key impediments to government carrying out the activities discussed so far: **Legal and ethical constraints**, and **public perception.** These barriers favour growing national SMA capacities outside government.

Legal and ethical constraints

Public servants have extensive obligations that constrain their behaviour, including but not limited to those described in the State Services Commission Model Standards, such as:

- **Propriety** – requirements to act in the public interest as public servants

- **Political neutrality** – requirements to both be and to appear politically neutral

- **Lawfulness and proportionality –** including rational connection to a legitimate purpose

- **Privacy** – requirements to maintain the anonymity of private citizens as much as possible

- **Algorithmic accountability** – including accountability for automated systems

- **Transparency** – disclosing collection and use of public information, compliance with Official Information and Public Records legislation

These obligations create practical barriers for government officials or contractors who wish to carry out effective capture and analysis of internet-based communications. For example:

- Monitoring of specific private individuals, as outlined in "a resource-efficient model for situational awareness" would almost certainly qualify as government surveillance. For government surveillance to take place on the basis of harmful speech, human rights law requires a substantial and specific case to be made in support: what specific communication, what kind of harm, to whom or to what, of what degree, what is the likelihood of harm, and even then, is the harm tolerable in a free and democratic society?

- Many of the most impactful forums for disinformation are not publicly accessible: e.g. closed Discord channels, WhatsApp groups, Telegram channels, or private Facebook groups. Gaining access to these typically requires that an investigator behave deceptively, e.g. by assuming a pseudonymous online identity. This bears resemblance to orthodox espionage tradecraft. The regulation of New Zealand government agencies and public servants strictly controls the circumstances and manner in which government officials may conduct this kind of activity.

- The Terms of Service for most social media platforms prohibit the large-scale 'scraping' of data without their express oversight and permission. While some jurisdictions have ruled that there are implicit exceptions for non-commercial research, the practice largely remains in a legal grey area. Despite this, many researchers will scrape data to prepare adequate datasets for study. While widely accepted in academia, this would be a risky practice for government researchers or contractors.

- Mass data collection by researchers often entails the advertent or inadvertent capture of personally identifying information, including names, phone numbers, addresses, and details of users' private lives. While some level of automated obfuscation of this information is standard in the field, this level is almost certainly insufficient to fully anonymise those whose communications are collected and could open the government to legal challenges.

- Disinformation actors – and those who knowingly or unknowingly spread their material – are often aware of and actively work to mitigate efforts to study and counter their efforts. The requirement for openness and transparency in government activity are likely to be abused by these groups to develop techniques to frustrate government SMA.

Public perception

Governments carrying out SMA will attract negative public attention, even if they do so lawfully and ethically. The practice touches on a host of public anxieties around privacy, free expression, and government influence over public discourse. Any significant government investment in SMA will produce narratives in the following vein:

- "The government is conducting surveillance against its political opponents"

- "The label of "disinformation" is being used to silence legitimate debate"

- "Government and social media companies are working together to control public opinion"

The traction and spread of these narratives can be mitigated somewhat by limiting the scope and enhancing the transparency of SMA efforts. Nevertheless, the spread and influence of these narratives will be difficult to predict and control, and will depend heavily on reactions by opposition parties, media, and civil society groups. These narratives can have substantive impacts, including:

- **Undermining trust in government** – both policy and personnel

- **Furthering radicalisation in fringe groups** that feel under threat

- **Driving audiences "off-platform"** to harder-to-access environments with more lax moderation and less visibility

- **Legitimating more extreme monitoring** practices by other states

- **Potential chilling of free speech** as people self-censor to avoid government observation – even those not taking part in mis- and disinformation

- **Discouraging cooperation** by useful partners, such as diplomatic partners, domestic and international civil society organisations and social media platforms

- **Delegitimising future efforts** to counter disinformation or regulate social media

No path forward alone

Ultimately, there are no good options for the New Zealand government as a lone actor in this space. The cutting edge of social media monitoring remains both time and resource intensive, and deeply imperfect. Platforms are struggling to meet even the standards they have set for themselves, despite access to all relevant data, full knowledge of their own systems, and access to leading experts. And government faces unique barriers to conducting effective social media analysis.

We also draw attention to the fact that recent reporting by RNZ and the New Zealand Herald has disclosed existing SMA efforts by DPMC which are expected to receive further investigation by the Office of the Privacy Commissioner.

Against this, we note the reality that:

- Members of the public have already called for enhanced social media monitoring.

- Agencies are already conducting some degree of monitoring, implying a perceived operational need for it.

- Public knowledge about the presence or absence of monitoring activities can play a deterrent effect toward external influence operations.

- The absence of monitoring may lead to unjustified assumptions that disinformation is occurring when it is not, undermining public trust unnecessarily.

- New Zealand may not detect disinformation activities which are occurring, meaning influence operations are successful in ways contrary to the public interest.

Questions around digital disinformation are only going to become more important, more complex, and more controversial in the future, especially as legislation in the European Union, the United Kingdom and other jurisdictions begins to be implemented. As a sovereign nation committed to multilateralism, human rights and the rule of law, New Zealand must prepare for that future in a way that amplifies our strengths and mitigates our limitations.

Released under the Official Information Act 1982

Part IV: Growing capacity outside government

Globally, there is a growing number of institutions, groups, and individuals outside government that are engaged in regular open-source analysis of internet communications. Groups like the Institute for Strategic Dialogue, the Election Integrity Partnership, and InterAction have produced world-leading research, conducted crucial outreach efforts, and provided valuable insights and advice to lawmakers.

Being wholly or partly outside of government helps these practitioners to produce useful analysis while maintaining public confidence. There is a strong case for supporting the development of an institution or group of this kind for the benefit of New Zealand – modelled on the best global examples, but reflective of New Zealand's unique cultural and legal characteristics. Beyond this, there is an opportunity for a hybrid non-State regulatory mechanism that entrenches relevant relationships between civil society groups, independent crown entities, and others.

The value of civil society groups

Even when the problem and its potential solutions are fully understood, the fight against disinformation and other social media harms will require a whole-of-society approach. There must be broad buy-in to the path taken – something which is unlikely if the work is perceived as a way for the government to exert influence over public discourse. The meaningful incorporation of a diverse array of voices on the issue will help counter fears of government overreach, facilitate the inclusion of key actors from the beginning of any further action, and allow the balancing of the many important perspectives and stakes inherent in SMA such as privacy, commercial considerations, and Te Tiriti obligations.

A civil society group is also likely to be a more appealing partner for social media platforms and other entities that are reluctant to directly collaborate with State governments and will allow New Zealand to position itself more effectively to take advantage of emerging transparency regimes under legislative initiatives, which grant greater data access to vetted researchers.

Why not simply support existing international organisations?

New Zealand has a number of unique features – among them a connection to mātauranga Māori, Te Tiriti obligations, comparatively high levels of social cohesion and media trust, and a highly specific socioeconomic milieu – that provide compelling reasons for undertaking New Zealand-based work rather than importing experts and conclusions from other jurisdictions. In addition, doing this work would grant New Zealand greater credibility in international engagements on the issue, and allow us to give better, more informed guidance to neighbours that may look to New Zealand for support as access to social media expands in the Pacific.

New Zealand has the opportunity to model best practices in addressing this important and highly charged issue in a way that is responsible, rigorous, and fully engaged with academia, civil society, indigenous perspectives, and the broader public.

Key design principles

While there are many factors that influence the ultimate success of any civil society group, Brainbox believes that any decisions should be made with seven principles in mind: **Full integration into civil society, data access as a priority, a cross-platform focus, continual self-assessment and development, explicability at all stages, insight from all sources, and Te Ao Māori centrality**. These are elucidated overleaf and owe a deep debt to those expressed in the ISD report "Developing a Civil Society Response to Online Manipulation".

Full integration into civil society

The priorities and direction of the proposed organisation should be informed by those of many groups and communities across civil society; it should work in ways that are transparent and understandable to civil society, and it should produce outputs that facilitate a civil societal response where possible. It must be able to work effectively with, and within, other civil society organisations and networks.

Data access as a priority

It must leverage the full opportunities available for civil society researchers to acquire data from all the platforms and online spaces relevant to illicit online manipulation, not just the largest and most easily accessible.

Cross-platform focus

It should conduct monitoring and research across a range of platforms. It is understood that disinformation efforts frequently occur across a number of platforms, often functionally separated for the purposes of planning, co-ordination and execution. Detections made on one platform may present either data collection opportunities on another platform, or input into the detection methodology on another platform.

Continual self-assessment and development

It must improve its own capabilities as it learns, identifying new techniques and developing the necessary tools to conduct effective investigations. Access to technological development capabilities will be essential, as will a cyclical structure that allows for past research to inform future planning, new conceptual frameworks, and innovative ways to respond to the challenges outlined in this report.

Explicability at all stages

This group should build confidence not only in its final outputs, but in the processes and systems used to generate them. As such, in addition to making sure its public-facing output is as clear and easy to understand as possible, it should have a visualisation and analysis function wherever the machine-driven parts of its detection system produce human-readable output or require manual intervention.

Insight from all sources

Rather than focusing entirely on detecting and analysing online information, it should draw insights from any area that may be useful; particularly traditional psychology, sociology, and representatives from communities that may be particularly vulnerable to disinformation campaigns.

Te Ao Māori centrality

In order to make sure Te Tiriti obligations are met, mātauranga Māori is respected, and the group maintains credibility with indigenous communities which tend to have lower confidence in government, Māori scholars and community representatives must be a core pillar of its structure from the beginning.

Selecting and attracting the right personnel is also crucial, and relationships with international practitioners should be encouraged. Our conversations with the global multidisciplinary community in this area were encouraging, and we expect that a mixture of local and overseas personnel is a realistic prospect. A prospective civil society group focused on SMA must have access, either through direct employment or wider networks, to the following expertise (drawing once again on ISD's report):

- **Data analysts and interrogators,** who can ensure the proper handling and analysis of data collected.

- **Visualisation,** technology and tool developers, who can represent discoveries accessibly and construct technological tools that the group needs.

- **Data journalists and subject matter experts,** who can understand and contextualise the analytical outputs from the system and identify the most promising leads for further investigation.

- **OSINT practitioners,** who can conduct targeted investigations of the most harmful, urgent and important detection that the system has made.

- **Legal experts in speech, platform law, and Te Tiriti,** who can ensure that work remains justifiable and within legal boundaries.

Structure and functions

We suggest two models worth investigating as a non-State approach to SMA in a way that is consistent with New Zealand's values:

1. Establish a civil society institution modelled on overseas institutions, such as DFR Lab or the Institute for Strategic Dialogue, which conducts SMA with a domestic focus at arm's length from government aside from funding support.

2. Establish a hybrid governance institution with an operational focus, that incorporates participation by Government as a stakeholder, alongside a range of other stakeholders that could also include Independent Crown Entities.

Full institutional design will take more detailed consultation. Regardless of the option chosen, we propose that the functions of any institution should include the following:

- **Conduct empirical work**: Fundamentally, the institution's mandate is to conduct empirical work, using SMA techniques. This work must be conducted in a way that allows appropriate scrutiny of its methods and techniques to build reliable knowledge about the online environment.

- **Publish outputs for operational use**: The institution must publish its findings in relatable and meaningful ways targeted to specific audiences. While there is some room for theoretical or meta-level discussions on relevant topics like definitions, outputs must be tailored toward its primary function: conducting SMA to build a meaningful picture of online communications for use in operational environments.

- **Maintain actual and perceived independence from government policy and influence**: the institution should be tasked with actively maintaining its perceived and actual independence from government policy.

- **Independent advocacy grounded in its empirical work**: the institution must have an advocacy and awareness-raising function. It will be critical for the institution to have an independent voice, particularly if it observes behaviour by States which is contrary to the law or the public interest. Importantly, this advocacy must be grounded in its empirical work in order to avoid straying into substantively political disputes that compromise its perceived independence.

- **Build broad global stakeholder relationships**: The institution ought to be tasked with building relationships with external institutions and research communities domestically and internationally. This would include key stakeholders such as governments, platforms, community organisations, academia, and others.

- **Build capacity to conduct high quality empirical work**: there are a range of training programmes being run by civil society institutions that teach people how to conduct safe, legal and ethical open source intelligence gathering and analysis. The institution ought to play a role in building capacity in New Zealand for conducting this kind of work, including by importing and exporting personnel, and upskilling New Zealanders with appropriate skill sets.

- **Direct advisory and commissioned investigations**: As a body with scarce expertise, the institution can provide direct advisory services to government and non-government actors. This could include being commissioned to conduct specific pieces of work, as well as using its expertise to tailor such work to reliable and meaningful outputs.

- **Explicit focus on human rights and a free and open internet, and accounting for New Zealand's specific socio-political context**: the institution must be mandated to support, promote and protect human rights and the preservation of a free, open and interoperable global internet. It must also be tasked with explicitly incorporating factors that make New Zealand what it is, including our values, culture and history.

Further considerations, such as recommended skills for a prospective oversight board and relationship to existing institutions, can be found in **Annex B: Institutional Considerations**.

Conclusion

SMA is already being undertaken by platforms, advertisers, researchers, and parts of government. However, as the field matures and states' approaches become more systematic, there is an absolute necessity to ensure that future SMA work is responsible, reliable, and ethical.

New Zealand has become a leading voice on social media issues in the wake of the Christchurch Call, and the approach we take to tackle disinformation will be replicated around the world. This proposal lays out a way to ensure that the values of the Christchurch Call are embedded in the way we understand and respond to emerging online communication issues, while protecting freedom of expression, privacy, and a plurality of voices in the public square.

Annex A: Case Studies

Example One: "Ill Advice: A Case Study in Facebook's Failure to Tackle COVID-19 Disinformation"
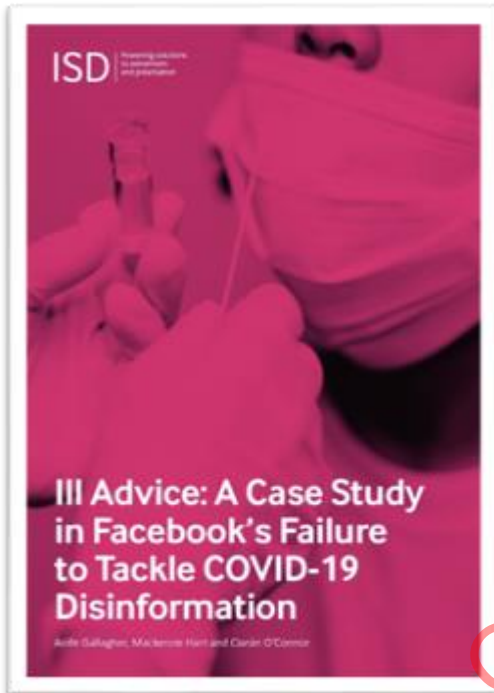
An ISD report on how effective Facebook (and to a lesser extent other social media platforms) have been in tackling Covid misinformation, primarily through the lens of a case study of a group called the "World Doctors Alliance" (WDA).

Their most crucial finding was that there's a lot of content on Facebook nearly (or wholly) identical to content that the platform has already taken down as misinformation. This indicates that Facebook's automated tools can't reliably identify content already flagged by the company as false.

The researchers used the CrowdTangle API to find Facebook posts mentioning the WDA or its members, then used an in-house data analytics tool to categorise them by language. They chose four languages to focus on (English, Spanish, German and Arabic), and then analysts manually analysed the 50 most popular posts in each language to determine whether they qualified as "disinformation" and confirm whether any platform action had been taken.

Key findings:
- 78% of the group's 1.2 million online followers are found on mainstream platforms (Facebook, Instagram, YouTube, Twitter, TikTok) which claim to prohibit vaccine misinformation.
- Large proportions – often the majority – of the most engaged-with content on Facebook mentioning the World Doctors Alliance or its members in English, Spanish, German and Arabic contained false, misleading or conspiratorial claims related to COVID-19 and vaccines.
- Organisations that are part of Facebook's factchecking program have debunked false claims made by the World Doctors Alliance 189 times since the beginning of the pandemic. Despite this extensive fact-checking effort, Facebook has not taken decisive action on the group or its members.
- ISD found minimal application of factchecking labels across the four languages analysed, with lower application rates on posts in German, Spanish and Arabic than in English. Content that does contain fact-checking labels was still accumulating tens and sometimes hundreds of thousands of engagements.
- Facebook failed to track down and label all versions of posts that have been deemed false by fact-checkers, despite claiming that they have AI technology that does this with a "very high degree of precision".
- Members of the World Doctors Alliance produce content in huge quantities. Facebook's one-at-a-time approach to fact-checking presents a huge challenge to fact-checkers and also allows the purveyors of disinformation to continue to spread false claims with little pushback.
- When information that is true (e.g. hospitals receive higher payments for COVID-19 patients) is used to spread a false narrative about the pandemic (e.g. case/death numbers are being manipulated), Facebook often does not label posts with additional context provided by fact-checkers.

Example Two: "The Long Fuse: Misinformation and the 2020 Election"
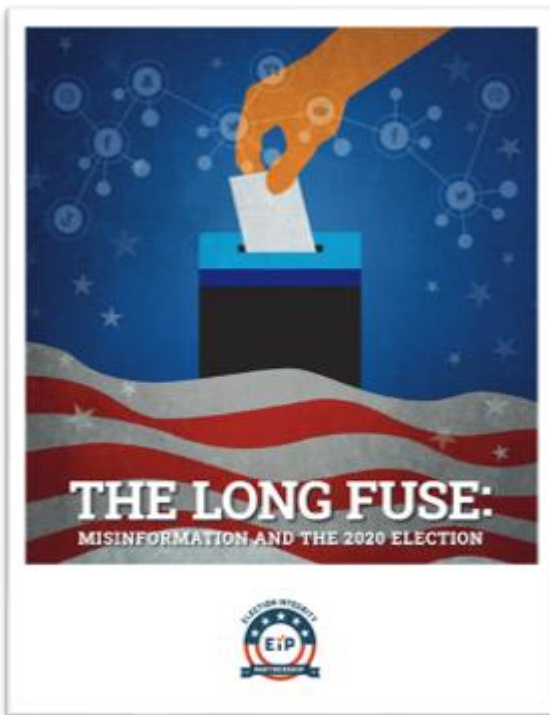
A report from the Election Integrity Partnership (EIP), a coalition of some of the foremost institutions in social media research and policy (The Stanford Internet Observatory, The University of Washington's Centre for an Informed Public, Graphika, and the Atlantic Council's Digital Forensic Research Lab) formed to combat voting related mis-and-disinformation in the 2020 US election season. Despite world-class expertise, tangible support from government, links to platforms, and a focused remit, the EIP arguably failed in its stated goal of countering election-related misinformation in the US; being unable to effectively combat widespread narratives discrediting the election's outcome, and failing to anticipate or prevent the January 6th Capitol riot.

The EIP had multiple tiers of on-call analysts and managers on shifts, taking in 'tickets': False or misleading claims flagged by in-house monitoring or partners in government, Civil Society, platforms, and news media. In total, they dealt with 639 "in-scope" tickets over the course of the project. The EIP relied primarily on external sources for fact-checking, and did not have a tasking relationship with these sources, limiting the scope they could cover.

The EIP's final report notes that their "per-ticket" analysis made it more difficult to identify and analyse overarching narratives. This cataloguing of narratives began only after the monitoring portion of their work had completed, sorting individual tickets post hoc.

Key findings:

- While the EIP and other researchers predicted a lot of the dynamics observed in practice, this did not translate into being able to prevent or combat them effectively.
- Lack of access to platform information made the task of external workers and researchers much harder.
- Non-falsifiable claims were a huge part of all narratives and very challenging for platforms.
- Framing was more impactful than individual pieces of information.
- There was a feedback loop between big media figures and grassroots movement, each generating narratives that would then be amplified by the other.
- There were networks of overlapping groups and audiences rapidly relaying pieces of misinformation.
- Cross-platform spread was the norm, with small, unregulated platforms like Parler generating some of the worst content.
- Each platform served a different purpose in the misinformation 'ecosystem' – for example, while Facebook was a place to reach large audiences and organize action, Twitter was a place to mobilize and "eventize" longer-form content stored elsewhere.
- Moderation was consistently inconsistent and lacked transparency, both hampering efforts to push back and inflaming the conspiracists further.

On moderation, the report notes that bad actors adapted quickly to changes made to platform policy and enforcement. Additionally, despite key "large spreader" accounts consistently exhibiting behaviour that should (enforcing the spirit and letter of the platform policies) have gotten them banned, platforms typically allowed them to remain up. This was sometimes justified with "newsworthiness" exceptions, but was often not justified at all.

The effects of "adding friction" to interactions with posts flagged as misleading were inconclusive, as were the true effects of content labelling – which the report notes was inconsistently and often incorrectly applied. There were also significant differences in which content got labelled depending on the platform. There are four problem areas that can't really be adequately addressed by platform policy in EIP's view: Cross-platform complexities, the use of non-falsifiable content, backlash against platform interventions, and organized outrage.

Example Three: "(Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures"

An academic paper studying the spread of misinformation on WhatsApp during the 2018 Brazilian election and national truck drivers' strike, producing valuable insights into the app's information flows. However, some of the techniques used arguably breach the platform's TOS, or would otherwise likely not be viable for government work.

Key methodological features:
- Researchers joined WhatsApp groups they found links to by searching Google and other social media sites manually. They did not proactively identify themselves as researchers.
- Though the groups were technically publicly accessible, most members likely had a reasonable expectation of privacy.
- They periodically downloaded all information from these groups and replaced telephone numbers and user names with unique identifiers. This anonymisation would not have removed other identifying data and would almost certainly not meet government privacy standards.
- The downloading of data, and what researchers did with it, arguably constitutes a breach of WhatsApp's TOS. WhatsApp has not objected, however.

- Researchers were able to automate the process of downloading images, reverse google searching them, finding if they'd appeared on fact checking sites, and extracting the sites' verdict without direct human involvement. While a clever solution, this system was capable only of identifying images already addressed by fact checkers.
- These researchers have created "WhatsApp monitor", which applies these sorts of techniques to a number of Brazilian and Indian WhatsApp groups and is still in use by researchers and reporters. This approach would be very likely do draw controversy if taken by the government.

Key findings:
- WhatsApp demonstrated similar network effects to more traditional social media platforms regarding the viral sharing and spread of content, despite limits on group size, due to crossover members between groups.
- Researchers claimed that 30% of captured images that were fact checked as misinformation could not be traced to prior sources, suggesting they were first posted on WhatsApp.
- WhatsApp was a very effective propagator of content to other sites – average time for a piece of content to be distributed beyond WhatsApp was around a week (less for unambiguous misinformation).
- A minority of groups were responsible for spreading the bulk of misinformation identified.

Annex B: Institutional Considerations

Relationship to existing institutions

There are some obvious institutions that may come to mind as existing institutional homes for these kinds of functions. We note some reservations about incorporating these functions into these existing institutions.

Academia and tertiary education institutions

Situating the institution within academia or an existing tertiary education institution initially has some appeal. In particular, such institutions perform empirical work frequently in situations of legal or ethical risk, and they are used to building connections across civil society. However, the following risks should be kept in mind:

- Generally speaking, academia and research institutions are not tasked with direct operational input. Their mandate is frequently to explore larger societal level issues which, in this case, are already well covered and may conflict with the institution's operational focus.

- Importantly, the institution we are proposing should itself be open to rigorous criticism given the nature of its activities and its functions, especially from academics and universities. There is a risk that situating the institution within an existing University or other tertiary or research institution might, in substance or perception, compromise the capacity of such institutions to act as critic and conscience in relation to the institution's functions.

He Whenua Taurikura – Centre for Countering Violent Extremism

We have considered whether this kind of function could sit with the recently implemented Centre for Countering Violent Extremism. In particular, there is at least a plausible relationship between disinformation and situations of radical violent extremism. In addition, there are likely to be areas of overlap when it comes to individual or community propensity to be radicalised by online communications, and the emerging literature on the role of platforms and algorithmic systems in contributing to this relationship. However, we have the following reservations about adding this function to He Whenua Taurikura, despite the high degree of potential overlap in subject matter between disinformation and countering violent extremism:

- Violent extremism by definition involves the adoption of violence as a legitimate political tool. It therefore justifies enhanced levels of state intervention, including the involvement of law enforcement and state surveillance. Linking the institution to this kind of use of state power may compromise its perceived independence, and enhance the perception that it is a tool of state surveillance and control.

- While there is some relationship suggested between disinformation or conspiracy theory content and violent extremism, in many cases mis- or disinformation will fall into a grey zone where the justification for both monitoring and intervention is much less clear cut. Put shortly, communications monitored on the basis they are "disinformation" are much more likely to be acceptable differences of political opinion with a range of plausibly pro-social intent, whereas that is seldom the case when it comes to communications being considered on the basis that they may be linked to violent extremism. On that basis, tasking a single institution with both monitoring disinformation and preventing violent extremism may lead to scope creep in a way that compromises the integrity of both programmes of work.

Key areas and skills for institution

Regardless of whether option 1 or option 2 is preferred, there should be an oversight board or committee for the institution. The members of that board should have demonstrable expertise in the following subject areas, noting that one member may be able to speak to multiple subject areas. The areas include:

- Te Tiriti and the requirements of Treaty partnership, including an understanding of New Zealand's colonial history

- Human rights and the rights of vulnerable or minority communities, including those protected by the prohibited grounds of discrimination in the Human Rights Act 1993

- Parliamentary democracy, rule of law, and constitutional government, with a specific focus on legality and the legal system, independent of any specific focus on platform regulation

- Expertise on theoretical aspects of propaganda and online information

- Expertise on empirical aspects of how communications impact human behaviour

- Expertise on digital technologies and their use for empirical purposes, including analysis of large data sets, and analysis of OSINT

- Connections to the international community, including international NGOs and human rights organisations

- While it may be inappropriate for the intelligence community to be directly represented, it will be important that the oversight board includes people with some understanding of intelligence gathering and national security frameworks, given the links with these subjects

- Expertise in the platform companies and an industry perspective

- Operational expertise in managing the volume and scale of complaints about online harms

- Expertise in Executive Government and the ability to bring a perspective from the needs of institutions like the Cabinet, as well as expertise in the operations and requirements of the public service

- Expertise in health, including public health and health systems, given the particular focus on the potential harms of health-related information

- Expertise in the non-governmental sector, including in community groups with a commitment to civil and human rights and limiting government over-reach

- Expertise in areas like geopolitics, international relations, diplomacy and international affairs

- Expertise in internet infrastructure and the requirements of and threats to a free and open internet

- Experience and expertise in governance of corporate entities, whether commercial or governmental or otherwise

In model 2, these areas could be covered by membership from Independent Crown Entities (such as the Human Rights Commission or the Office of the Privacy Commission), and one or two representatives of an appropriate Crown Agency, such as DPMC, or the Ministry of Health.

Released under the Official Information Act 1982

Bibliography

"Anti-Lockdown Activity: Germany Country Profile." ISD. Accessed May 26, 2022. https://www.isdglobal.org/isd-publications/anti-lockdown-activity-germany-country-profile/.

"Anti-Muslim Hate." CCDH. Accessed April 29, 2022. https://www.counterhate.com/anti-muslim-hate.

Appelman, Naomi, Stephan Dreyer, Pranav Bidare, and Keno Potthast. "Truth, Intention and Harm: Conceptual Challenges for Disinformation-Targeted Governance." Internet Policy Review. Accessed May 23, 2022. https://policyreview.info/articles/news/truth-intention-and-harm-conceptual-challenges-disinformation-targeted-governance/1668.

"Between Conspiracy and Extremism: A Long COVID Threat? An Introductory Paper." ISD. Accessed May 26, 2022. https://www.isdglobal.org/isd-publications/between-conspiracy-and-extremism-a-long-covid-threat-introductory-paper/.

Bradford, Ben, Florian Grisel, Tracey Meares, Emily Owens, Baron Pineda, Jacob Shapiro, Tom Tyler, and Danieli Peterman. "Report Of The Facebook Data Transparency Advisory Group." Justice Collaboratory. Accessed June 9, 2022. https://www.justicehappenshere.yale.edu/reports/report-of-the-facebook-data-transparency-advisory-group.

Burton, Jason W., Nicole Cruz, and Ulrike Hahn. "Reconsidering Evidence of Moral Contagion in Online Social Networks." Nature Human Behaviour 5, no. 12 (December 2021): 1629–35. https://doi.org/10.1038/s41562-021-01133-5.

Cameron, Dell, Shoshana Wodinsky, and Mack DeGuerin. "We're Publishing the Facebook Papers. Here's What They Say About Donald Trump, the 2020 Election, and Jan. 6." Gizmodo, April 18, 2022. https://gizmodo.com/facebook-papers-donald-trump-2020-election-jan-6-capito-1848698220.

Carl Miller [@carljackmiller]. "The World of Disinformation Response Is Growing Fast. Lots of New Commercial Offerings. For Anyone Looking at These, There Are a Few Things to Avoid: - Any Detection Capability Predicated on Some Secret, Proprietary, Universally-Applicable Algorithm. Full Stop. 1/." Tweet. Twitter, April 21, 2022. https://twitter.com/carljackmiller/status/1517082933047271424.

Claire Wardle, Hossein Derakhshan. "Information Disorder: Toward and Inderdisciplinary Framework for Research and Policy Making," 2017. https://rm.coe.int/information-disorder-report-version-august-2018/16808c9c77.

Clegg, Nick. "Combating COVID-19 Misinformation Across Our Apps." Meta (blog), March 25, 2020. https://about.fb.com/news/2020/03/combating-covid-19-misinformation/.

Clothier, Brent. "He Pānui Statement." Royal Society Te Apārangi. Accessed May 23, 2022. https://www.royalsociety.org.nz/news/he-panui-statement/.

Colliver, Chloe, and Carl Miller. "Developing a Civil Society Response to Online Manipulation." ISD, n.d. https://www.isdglobal.org/isd-publications/developing-a-civil-society-response-to-online-manipulation/.

Comerford, Milo, Jakob Guhl, and Carl Miller. "Understanding the New Zealand Online Extremist Ecosystem." ISD, n.d.

"COVID-19 Misinformation - Twitter Report." Twitter. Accessed June 9, 2022. https://transparency.twitter.com/en/reports/covid19.html.

Deutsche Akademie Der Naturforscher Leopoldina, Deutsche Akademie Der Technikwissenschaften, and Union Der Deutschen Akademien Der Wissenschaften. "Digitalisation and Democracy." Series on Science-Based Policy Advice: Position Paper. MyCoRe Community, 2021. https://doi.org/10.26164/LEOPOLDINA_03_00407.

@DFRLab. "Polish-Language Telegram Channels Spread Anti-Refugee Narratives." DFRLab (blog), May 31, 2022. https://medium.com/dfrlab/polish-language-telegram-channels-spread-anti-refugee-narratives-aaf3ffdc81ed.

Wardle, Claire. "Information Disorder, Part 1: The Essential Glossary." First Draft Footnotes, July 9, 2018. https://medium.com/1st-draft/information-disorder-part-1-the-essential-glossary-19953c544fe3.

Dreyer, Stephan, Pranav Bidare, and Clara Keller. "Between Evidence and Policy: Bridging the Gap in Disinformation Regulation." Internet Policy Review. Accessed May 30, 2022. https://policyreview.info/articles/news/between-evidence-and-policy-bridging-gap-disinformation-regulation/1667.

Epstein, Ben. "Why It Is So Difficult to Regulate Disinformation Online." In The Disinformation Age: Politics, Technology, and Disruptive Communication in the United States, edited by Steven Livingston and W. Lance Bennett, 190–210. SSRC Anxieties of Democracy. Cambridge: Cambridge University Press, 2020. https://doi.org/10.1017/9781108914628.008.

Farajtabar, Mehrdad, Jiachen Yang, Xiaojing Ye, Huan Xu, Rakshit Trivedi, Elias Khalil, Shuang Li, Le Song, and Hongyuan Zha. "Fake News Mitigation via Point Process Based Intervention," n.d., 10.

Fathaigh, Ronan Ó, Natali Helberger, and Naomi Appelman. "The Perils of Legally Defining Disinformation." Internet Policy Review 10, no. 4 (November 4, 2021). https://policyreview.info/articles/analysis/perils-legally-defining-disinformation.

Funke, Daniel, and Daniela Flamini. "A Guide to Anti-Misinformation Actions around the World." Poynter (blog). Accessed May 30, 2022. https://www.poynter.org/ifcn/anti-misinformation-actions/.

Gallagher, Aoife, Mackenzie Hart, and Ciarán O'Connor. "Ill Advice: A Case Study in Facebook's Failure to Tackle COVID-19 Disinformation." ISD, n.d.

Haidt, Jonathan, and Chris Bail. "Social Media and Political Dysfunction." Google Docs. Accessed May 30, 2022. https://docs.google.com/document/u/0/d/1vVAtMCQnz8WVxtSNQev_e1cGmY9rnY96ecYuAj6C548/edit?usp=embed_facebook.

Hall, Kristin. "Misinformation: Down the Rabbit Hole, and Back." 1 News. Accessed May 26, 2022. https://www.1news.co.nz/2022/04/04/misinformation-down-the-rabbit-hole-and-back/.

Hameleers, Michael, Edda Humprecht, Judith Möller, and Jula Lühring. "Degrees of Deception: The Effects of Different Types of COVID-19 Misinformation and the Effectiveness of Corrective Information in Crisis Times." Information, Communication & Society 0, no. 0 (December 31, 2021): 1–17. https://doi.org/10.1080/1369118X.2021.2021270.

Hannah, Kate, Sanjana Hattotuwa, and Kayli Taylor. "Working Paper: The Murmuration of Information Disorders," 2022, 22.

Harvey, David. "Fear Itself?" The IT Countrey Justice (blog), May 20, 2022. https://theitcountreyjustice.wordpress.com/2022/05/20/fear-itself/.

Global Witness. "How Facebook's Algorithm Amplifies Climate Disinformation." Accessed April 14, 2022. https://www.globalwitness.org/en/campaigns/digital-threats/climate-divide-how-facebooks-algorithm-amplifies-climate-disinformation/.

"Information Operations - Twitter Report." Twitter. Accessed June 9, 2022. https://transparency.twitter.com/en/reports/information-operations.html.

Ingram, Matthew. "Facebook 'Transparency Report' Turns out to Be Anything But." Columbia Journalism Review. Accessed June 9, 2022. https://www.cjr.org/the_media_today/facebook-transparency-report-turns-out-to-be-anything-but.php.

Atlantic Council. "Inside a New Effort to Define and Promote Tech Transparency," December 14, 2021. https://www.atlanticcouncil.org/news/transcripts/inside-a-new-effort-to-define-and-promote-tech-transparency/.

Jack, Caroline. "Lexicon of Lies." Data & Society. Data & Society Research Institute, August 9, 2017. https://datasociety.net/library/lexicon-of-lies/.

"January 2022 Coordinated Inauthentic Behavior Report." Meta, February 16, 2022. https://about.fb.com/news/2022/02/january-2022-coordinated-inauthentic-behavior-report/.

Jungherr, Andreas, and Ralph Schroeder. "Disinformation and the Structural Transformations of the Public Arena: Addressing the Actual Challenges to Democracy." Social Media + Society 7, no. 1 (January 1, 2021): 2056305121988928. https://doi.org/10.1177/2056305121988928.

Keller, Clara Iglesias. "Don't Shoot the Message: Regulating Disinformation Beyond Content." Direito Público 18, no. 99 (October 28, 2021). https://doi.org/10.11117/rdp.v18i99.6057.

Keller, Daphne. "Some Humility About Transparency." Accessed April 14, 2022. https://cyberlaw.stanford.edu/blog/2021/03/some-humility-about-transparency.

Keller, Daphne, and Paddy Leerssen. "Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation." In Social Media and Democracy, edited by Nathaniel Persily and Joshua A. Tucker, 1st ed., 220–51. Cambridge University Press, 2020. https://doi.org/10.1017/9781108890960.011.

Lapowsky, Issie. "Why Facebook's Data-Sharing Project Ballooned into a 2-Year Debacle." Protocol, February 14, 2020. https://www.protocol.com/facebook-data-sharing-researchers.

Lewandowsky, Stephan, and Sander van der Linden. "Countering Misinformation and Fake News Through Inoculation and Prebunking." European Review of Social Psychology 32, no. 2 (July 3, 2021): 348–84. https://doi.org/10.1080/10463283.2021.1876983.

Lwin, May Oo, Jiahui Lu, Anita Sheldenkar, Peter Johannes Schulz, Wonsun Shin, Raj Gupta, and Yinping Yang. "Global Sentiments Surrounding the COVID-19 Pandemic on Twitter: Analysis of Twitter Trends." JMIR Public Health and Surveillance 6, no. 2 (May 22, 2020): e19447. https://doi.org/10.2196/19447.

Lyons, Tessa. "Seeing the Truth." Meta (blog), September 13, 2018. https://about.fb.com/news/2018/09/inside-feed-tessa-lyons-photos-videos/.

MacCarthy, Mark. "Transparency Recommendations for Regulatory Regimes of Digital Platforms." CIGI, 2022, 22.

Mcclain, Colleen, Regina Widjaya, Gonzalo Rivero, and Aaron Smith. "The Behaviors and Attitudes of U.S. Adults on Twitter." Pew Research Center: Internet, Science & Tech (blog), November 15, 2021. https://www.pewresearch.org/internet/2021/11/15/the-behaviors-and-attitudes-of-u-s-adults-on-twitter/.

Michael Bang Petersen [@M_B_Petersen]. "Today, I Stood before the Danish Parliament on a Public Hearing on Social Media & Democracy As a Researcher of Online Hate, I Could Have Spent Hours. But I Had 10 Minutes, so I Had to Be Focused The Title Was 'The Myths About Social Media' Here Is What I Said 🗣 📝 (1/12)." Tweet. Twitter, January 18, 2022. https://twitter.com/M_B_Petersen/status/1483457679800651787.

Miller, Carl, Jakob Guhl, and Milo Comerford. "ISD NZ Report: Methodological Discussion." ISD, n.d.

Mohan, Neal. "Inside Responsibility: What's next on Our Misinfo Efforts." blog.youtube. Accessed June 9, 2022. https://blog.youtube/inside-youtube/inside-responsibility-whats-next-on-our-misinfo-efforts/.

Morgan, Kevin. "Taking Action against COVID-19 Vaccine Misinformation." Newsroom | TikTok (blog), August 16, 2019. https://newsroom.tiktok.com/en-gb/taking-action-against-covid-19-vaccine-misinformation.

Nimmo, Ben. "The Breakout Scale: Measuring the Impact of Influence Operations." Brookings Institute, n.d.

Nimmo, Ben, David Agranovich, and Nathaniel Gleicher. "Adversarial Threat Report #1 (Q1 2022)." eta, n.d.

O'Connor, Ciarán. "The Conspiracy Consortium: Examining Discussions of COVID-19 Among Right-Wing Extremist Telegram Channels." ISD, n.d.

Odabaş, Meltem. "10 Facts about Americans and Twitter." Pew Research Center (blog). Accessed May 30, 2022. https://www.pewresearch.org/fact-tank/2022/05/05/10-facts-about-americans-and-twitter/.

OEDC Digital Economy Papers. "Transparency Reporting on Terrorist and Violent Extremist Content Online : An Update on the Global Top 50 Content Sharing Services | En | OECD." Accessed April 27, 2022. https://www.oecd.org/digital/transparency-reporting-on-terrorist-and-violent-extremist-content-online-8af4ab29-en.htm.

Park, Sora, Kerry McCallum, Jee Young Lee, Kate Holland, Kieran McGuinness, Caroline Fisher, and Emma John. "COVID-19: Australian News and Misinformation Longitudinal Study." Report. News and Media Research Centre, March 21, 2022. Australia. https://apo.org.au/node/316582.

Persily, Nathaniel, and Joshua A. Tucker, eds. Social Media and Democracy: The State of the Field, Prospects for Reform. 1st ed. Cambridge University Press, 2020. https://doi.org/10.1017/9781108890960.

Pickles, Kristen, Erin Cvejic, Brooke Nickel, Tessa Copp, Carissa Bonner, Julie Leask, Julie Ayre, et al. "COVID-19 Misinformation Trends in Australia: Prospective Longitudinal National Survey." Journal of Medical Internet Research 23, no. 1 (January 7, 2021): e23805. https://doi.org/10.2196/23805.

"Platform Manipulation - Twitter Report." Twitter. Accessed June 9, 2022. https://transparency.twitter.com/en/reports/platform-manipulation.html.

Resende, Gustavo, Philipe Melo, Hugo Sousa, Johnnatan Messias, Marisa Vasconcelos, Jussara Almeida, and Fabrício Benevenuto. "(Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures." In The World Wide Web Conference, 818–28. San Francisco CA USA: ACM, 2019. https://doi.org/10.1145/3308558.3313688.

Rosen. "An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19." Meta (blog), April 16, 2020. https://about.fb.com/news/2020/04/covid-19-misinfo-update/.

Rossini, Patrícia, Jennifer Stromer-Galley, Erica Anita Baptista, and Vanessa Veiga de Oliveira. "Dysfunctional Information Sharing on WhatsApp and Facebook: The Role of Political Talk, Cross-Cutting Exposure and Social Corrections." New Media & Society 23, no. 8 (August 1, 2021): 2430–51. https://doi.org/10.1177/1461444820928059.

"Rules Enforcement - Twitter Report." Twitter. Accessed June 9, 2022. https://transparency.twitter.com/en/reports/rules-enforcement.html.

Sarang. "Community Standards Enforcement Report Assessment Results." Meta (blog), May 17, 2022. https://about.fb.com/news/2022/05/community-standards-enforcement-report-assessment-results/.

Schulz, Wolfgang, Stephan Dreyer, Elena Stanicu, and Keno Potthast. "Disinformation: Risks, Regulatory Gaps and Adequate Countermeasures. Expert Opinion Commissioned by the Landesanstalt Für Medien NRW." Leibniz Institute for Media Research | Hans-Bredow-Institut, November 10, 2021.

Scott, Hamilton. "As a Matter of Fact." North & South Magazine, April 3, 2022. https://northandsouth.co.nz/2022/04/03/richard-dawkins-matauranga-maori-debate/.

Shu, Kai, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. "Fake News Detection on Social Media: A Data Mining Perspective." arXiv, September 2, 2017. http://arxiv.org/abs/1708.01967.

BRAINBOX

Auckland / Wellington
New Zealand

info@brainbox.institute
www.brainbox.institute