Under embargo until 10.30am on Tuesday 9 November

Working Paper: Mis- and disinformation in Aotearoa New Zealand from 17 August to 5 November 2021

The Disinformation Project

Kate Hannah, Sanjana Hattotuwa, Kayli Taylor

Te Pūnaha Matatini; Department of Physics, University of Auckland; Centre for Science in Society, Te Herenga Waka.

Introduction

Since February 2020 a small interdisciplinary team, The Disinformation Project, has been observing and analysing open source publicly available data related to Covid-19 mis- and disinformation on social media, mainstream media, and in physical and other digital forms of information and knowledge dissemination. Our project is part of the Aotearoa New Zealand National Centre of Research Excellence for Complexity, Te Pūnaha Matatini. In our work, The Disinformation Project has developed a novel mixed methods approach which combines a range of standard open-source quantitative reporting from social media, media platforms or sources with a rich text and artefact-based narrative analysis of longform qualitative data. From August 2020, our work included the study of mis- and disinformation ecosystems in Aotearoa, including the seed and spread of ‘dangerous speech’, hateful expression, and criminal behaviour. We focus on effects and causes here, but study the global trends, themes, narratives, and actors who influence online harms in Aotearoa.

Our novel approach embraces quantitative measures based on the volume, vectors and velocity of inaccurate content, amplification of mis- and disinformation by groups and individuals, tracking of narratives across online and offline contexts and key distribution signatures. Furthermore, and significantly, harmful content is also qualitatively analysed through gendered, country-specific, and other contextual frames. The reporting is presented in ways which are immediately usable for decision-makers, alongside media commentary on the harms that mis- and disinformation and ‘dangerous speech’ present to social cohesion, freedom of expression, inclusion, and safety.

Since 17 August 2021, when Aotearoa New Zealand’s Delta outbreak meant a shift into Covid-19 Alert Level 4 across the country, there was a sharp increase in the popularity and intensity of Covid-19 specific disinformation and other forms of ‘dangerous speech’ and disinformation, related to far-right ideologies. Over the past twelve weeks, The Disinformation Project monitored this material, observing key trends, and analysing impact. This brief working paper introduces some of our key findings so far on the infodemic – around engagement; content, reception to the Covid-19 vaccine, language, approaches employed and targeted groups.

Definitions

Mis- and disinformation is transmitted within and across platforms, and often very rapidly reaching audiences in the tens of thousands. The Disinformation Project describes these complex phenomena as "ecologies" – systems and networks that mirror and migrate content, discourses, language, and values across different platforms to audiences, with significant online impact and growing offline consequences. For the purposes of our study we use the following definitions from Berentson-Shaw and Elliot:

- 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"[1]

When we describe conspiratorial thought, we draw from Dentith's work for a simple definition of conspiracy theory, defining them as purported explanations which cite a conspiracy at the salient cause of some event or phenomenon.[2] The Disinformation Project uses the category 'dangerous speech' to classify material observed and analysed with these ecologies. Susan Benesch's Dangerous Speech Project asserts that: "dangerous speech is any form of expression (e.g. speech, text or images) that can increase the risk that its audience will condone or participate in violence against members of another group."[3] The hallmarks of 'dangerous speech' are useful analytical tools for our work also.

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.

The most recent Covid-19 outbreak and the vaccination are highly visible, potent symbols used to push various far-right and conservative ideologies around issues such as gun control, rural land rights and 1080, Māori sovereignty and water/land rights, 'free

[1] Berentson-Shaw J and Elliot M. *Misinformation and Covid-19: a briefing for media*. Wellington: The Workshop; (2020).
[2] 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).
[3] The Dangerous Speech Project*, Dangerous Speech: A Practical Guide*: 19 April 2021 https://dangerousspeech.org/guide/

speech', faith (Christian evangelical or Pentecostal), abortion, euthanasia, cannabis law reform, families and family structure, LGBTQIA+ rights, including conversion therapy, immigration, race, and gender. The growing polarisation, engineered by leading mis- and disinformation producers within Aotearoa New Zealand, between those who are vaccinated and those who are not, seeks to normalise the increasingly intense negotiation of difference.

Posts and engagement

The Disinformation Project observes a large number of publicly available groups, pages, and accounts within Aotearoa's disinformation ecology. The platforms we observe include, but are not limited to: Telegram, Facebook Pages, Facebook Groups, Facebook accounts, Instagram, Twitter and any sign-posted, off-platform content harbours, like the .nz top-level domain, other websites and platforms like Rumble, Odysee, Gab, and Gettr. Based on the grounded, daily analysis of very large volumes of data, The Disinformation Project reports on emergent trends, themes and signals within a disinformation landscape that is sophisticated, motivated, adaptive, resilient, increasingly violent and significantly volatile.

Both posts and engagement have drastically increased since 17 August 2021 and show a trajectory of growth and spread that is increasing, widening, and deepening every week. Disinformation activity shows a high degree of coordination and collaboration within and between platforms. This means that content shared into one platform is quickly shared amongst other accounts within the same platform or app, and also transmits across different social media platforms. Since mid-August, Telegram emerged as the platform of choice for the spread of mis- and disinformation in Aotearoa New Zealand. As a platform, Telegram does not feature oversight of, or policies around mis- and disinformation.

The volume of content studied by The Disinformation Project since 17 August, across all platforms, is significant and cumulatively in the hundreds of millions of data points. Quantitative analysis around the mis- and disinformation volume (amount of content), vectors (platforms and apps content is produced and shared on) and velocity (speed at which content is produced) since mid-August, under the Delta Level 4 lockdowns, is unprecedented. We note that it is by order of magnitude more than the content seed and spread over 2020, and even in the first half of 2021.

The Disinformation Project also studied, and for the first time in Aotearoa New Zealand, a circadian rhythm associated with mis- and disinformation content production across the public Facebook Pages, Groups, Instagram accounts and Telegram channels studied. Our research clearly flags the degree to which there is pattern to each day's production of content, with peaks in the morning, afternoon and evening. These peaks drive engagement throughout the day, and for a longer period each day. The instrumentalisation of anxiety, through performative outrage, sensationalist content, conspiratorialism and the associated phenomenon of apophenia, combined with complicated algorithmic and network effects, is as much a problem online as it is offline, influencing psycho-social well-being and mental health.

The study of Aotearoa New Zealand's mis- and disinformation ecologies mirrors the work by the Centre for Countering Digital Hate (CCDH), and their consequential Disinformation Dozen report.[4] A few accounts, increasingly, generate the most amount of mis- and disinformation, which is shared far more widely. These accounts are increasingly coordinated in the production of content, and the selection of frames, subjects, issues, topics and offline events. The Disinformation Project's work reveals complex network effects as a consequence of this strategic, sophisticated production of harmful content by a few highly motivated actors, that tens of thousands of others go on to share.

Content

Reception to the Covid-19 vaccine

We have observed a critical shift from vaccine hesitancy to vaccine resistance within the core groups we study on Telegram. Here, with no platform-level guidelines or interventions such as the use of interstitials which may be present on other, more mainstream social media platforms, the discourse has shifted over the period of this study (from 17 August) from hesitancy and uncertainty to one of active resistance and refusal. Counter speech in an environment of active resistance, where individuals have attached identity to refusal, is difficult and in fact most often leads to further entrenchment of identity within this framework. Telegram channels and groups proliferate content which is violent, far-right, and related to the conspiracy theory QAnon, signalling a near-frictionless shifting of New Zealanders from vaccine hesitancy, to vaccine resistance, and then to content reflective of wider conspiratorial ideologies.

Approaches

Memes/memetic material is highly popular across the platforms. So too is the use of humour and trolling, as it is deemed psychologically more effective. Memetic material pre-frames content for easy sharing. Common themes include encouraging people to "take a stand" on one's social media with vaccination messages, offer increasing identity-based divisiveness, and ridicule public health guidelines. The in-group jokes offered by memetic content offers strong in-group identity markers and provides opportunity to make fun of out-group members – either the vaccinated or the state. This separation into in-group and out-group cements difference and encourages division. This content ranges from humorous re-purposing of known memes to explicit sexist and racist content, imagery of death and execution, and historic imagery from the Holocaust, the Cold War, and other violent or extremist events. Nazi references, written,

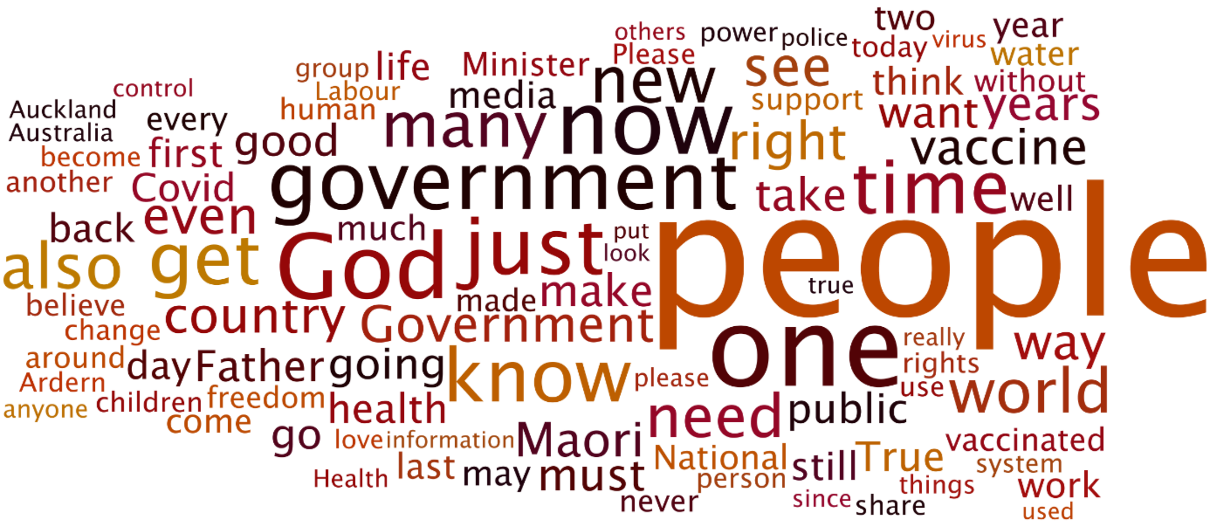
[4] Center for Countering Digital Hate. "The Disinformation Dozen" https://www.counterhate.com/disinformationdozen (2021)

spoken or visually expressed, are normative in many of these online ecologies studied, and have become increasingly so since 17 August 2021.

In addition to memetic content, testimonies of the unvaccinated remain a form of content with high potential for virality. One cluster studied targets both the currently unvaccinated and those who have received one but not both doses of the vaccine. These testimonials have emotive potential within these populations. Such emotive testimonials are likely to be believed and sympathised with – and should be considered as presenting genuine harm to ongoing efforts to increase vaccine uptake. Emotional, and often disturbing content provided via testimonies tends to focus on vaccine harms experienced by women and girl children – a known hallmark of 'dangerous speech' is the presentation of harm or attack on women and girls.

Other content tends to focus on a cluster of notions related to state versus individual rights and responsibilities. It is here that international content, particularly from the US and Australia, is repackaged and re-shared in Aotearoa New Zealand, framing a state that is described as tyrannical, treacherous, and forsaking international and/or national law. These frames are where some of the most extreme content is observed: violent imagery and language, threats to groups and individuals, allusions to historic violent events and war or crime. In the last month, we have observed more content which connects events in Aotearoa New Zealand with the Q conspiracy, and with far-right conspiratorial narratives more generally. These include white supremacist, Incel or extreme misogyny, Islamophobia and anti-migrant sentiment, and anti-Semitism. We have also observed increasing levels of anti-Māori racism.

Language



This wordcloud of text from Facebook Group posts in August and September 2021 – showing a focus on 'people' who are invited to resist, call out, laugh at, and ridicule public health measures, vaccinations, and the risks of Delta. 'Government' is also a target, and there is a high number of Christian-related tropes and framing. The posts show that resistance to vaccines is being framed as something that is Christian and faith-based. Here we observe that while vaccine and Covid-19 remain the purported purpose of these groups, other motivations are becoming clearer – here within this cluster of groups, a conservative, sometimes faith-based identarian agenda which focuses on traditional family structures and gender roles.

Language and word/image choice have shifted during this period of our study. The last twelve weeks showing increasingly violent language and other forms of expression, which has become normalised and justified within the groups and individuals who make up the disinformation community in-group. Language specifically targeting individuals and minority groups has become more violent and graphic. This shift has been marked in the way in which discourse, symbols and memes we would expect to observe on fringe social media platforms has made its way to mainstream social media, and mainstream media-driven conversations. This new normal includes explicit terminology, violent jokes, transphobia and homophobia, casual racist invective and slurs, crudity and vulgarity – and in use regularly by a wide range of New Zealanders.

The limitations of 'dangerous speech'

The level of engagement and changes in the type of content produced over the last twelve weeks have led The Disinformation Project to re-define how we describe and analyse 'dangerous speech'. 'Dangerous speech' is defined as any form of expression (e.g., speech, text, or images) that can increase the risk that its audience will condone or commit violence against members of another group.[5] This description remains central to our approach, but this useful definition comes with a number of caveats set out by The Dangerous Speech Project. Firstly, this definition confines violence to "direct physical or bodily harm" and explicitly excludes doxing, incitement to self-harm, or discrimination. Many other definitions of violence include non-physical harm: Johan Galtung described structural violence as including discrimination, exclusion and exploitation,[6] while the UN Declaration of the Elimination of Violence Against Women defines violence against women as "gender-based violence that results in, or is likely to result in, physical, sexual or psychological harm or suffering to women."[7]

The notion of slow violence, grounded in justice movements describes how "the temporal dispersion of slow violence impacts the way we perceive and respond to a

[5] The Dangerous Speech Project. *Dangerous Speech: A Practical Guide*: 19 April 2021 https://dangerousspeech.org/guide/
[6] Galtung, Johan. "Violence, Peace, and Peace Research", *Journal of Peace Research* 6, no. 3 (1969): 167-191.
[7] United Nations. *Declaration on the Elimination of Violence against Women,* (1993).

variety of social afflictions – from domestic abuse to post-traumatic stress – but has especially powerful implications for environmental calamities”.[8]

Benesch (and the Dangerous Speech Project’s) definition of ‘dangerous speech’ has other limitations. Conceived within a context of political violence with clearly or easily defined in-groups and out-groups, the examples given by the project itself in defining how speech targeting individuals is outside the scope of ‘dangerous speech’ are telling. In Aotearoa New Zealand, women associated with the Covid-19 response as politicians, healthcare professionals or experts are targeting individually for harassment, including non-consensual video recordings, but also framed as representative of transgressive women who are then targeted with highly misogynistic framing, including death and rape threats. Operationally, The Disinformation Project draws on the definition of ‘dangerous speech’, and its hallmarks:

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

These above hallmarks are highly useful for the purposes of defining speech which is likely to condone or encourage committing violence. But we note that since Benesch’s original conceptualisation[9] the Internet and related web technologies, including social media platforms, have significantly changed group identity dynamics, the impact of which is repeatedly noted in counter-terrorism studies.[10] Conventional definitions of political terrorism are markedly different, and so-called ‘lone wolf actors’ are likely to have been radicalised in a manner akin to slow violence, raised within contexts of widespread and normalised misogyny and white supremacist and other racist thought amongst other ideologies and narrative frames prevalent on social media and in others media discourses.

So while we use the definition, the hallmarks, and the wider framework – the message itself, the audience, the historical and social context of the message, the speaker, and the medium used – we have, in response to the proliferation of ‘dangerous speech’ content present in Aotearoa New Zealand’s mis- and disinformation ecosystem, expanded the definition to consider violence as articulated against individuals as

[8] Nixon, Rob. *Slow Violence, Gender, and the Environmentalism of the Poor*, (2011).
[9] Benesch, Susan. “Vile Crime or Inalienable Right: Defining Incitement to Genocide”, *Virginia Journal of International Law,* (2003).
[10] Winter, Charlie et al. “Online Extremism: Research Trends in Internet Activism, Radicalization, and Counter-Strategies”, International Journal of Conflict and Violence, (2020).

representative of groups, particularly in the case of clearly gendered or racialised 'dangerous speech'.

In this manner, given we are taking an ecosystems approach to understanding the interconnected networks of disinformation and 'dangerous speech', we view gendered and racialised speech as akin to slow violence – an 'environmental' calamity which is most often rendered ***invisible*** to *those who do not experience its harms*. Such harms are exacerbated by algorithmic harms, which are not neutral in creation or outcome, but instead reinforce oppressive social relationships.[11] Therefore, instead of accepting the standard definition that slow violence occurs 'out of sight', we must instead ask the question: 'out of sight to whom?'.[12]

Targeted groups and individuals

The online mis- and disinformation content The Disinformation Project observes targets key individuals and groups. This includes but is not limited to Māori, Pasifika, migrants and ethnic minorities, women, gender minorities, LGBTQIA+ people, people with disabilities, health workers, government employees. In addition, leading public figures and officials– including members of Parliament, journalists, health officials, academics, and community leaders – receive specific targeting and abuse.

For example, mainstream media's reporting on the uptake of vaccination by Māori has increased a perception of Māori as vaccine hesitant and anti-vaccination, which has been picked up within circles of disinformation in way that capitalises on racism and further targets disinformation towards those groups. This allows for the targeting of Māori, and the intensification of anti-Māori racism within mis- and disinformation circles.

The 'hīkoi' that took place on Tuesday 26 October – Wednesday 27 October 2021 relied on Māori motifs and symbols, such as the language of 'hīkoi' and the United Tribes flag. A large number of individuals who appear cis-male and Pākehā dominated Telegram and Facebook content about the hīkoi. This aligns with an increasing use of Māori voices, narratives, and imagery for agendas of white supremacist individuals and groups who make up one cluster we study.

Cherry-picked medicinal and scientific data is shared to support vaccine hesitancy. This particularly targets people who menstruate and pregnant people. Under researched vaccine side effects, such as its effect on menstrual cycles, have been used to targeted people who menstruate with mis- and disinformation. For people undergoing IVF and/or trying to conceive, these are legitimate concerns – which have been exploited by local

[11] Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism,* New York: NYU Press, (2018).
[12] Davies, Thom. "Slow violence and toxic geographies: 'Out of sight' to whom?", *Politics and Space,* (2019).

groups attempting to push Covid-19 related mis-and disinformation. One cluster in particular continues to rely on fears experienced by pregnant people around vaccination to fuel distrust. It also uses this to push its values – including transphobia, ideas about motherhood as sainthood, and traditional gender roles. Mis- and disinformation and conspiratorial thinking is actively targeting those who menstruate – particularly those who are pregnant or trying to conceive.

Public figures – particularly minorities – continue to receive disproportionate harmful attention. Figures such as the Prime Minister receive a particularly insidious form of abuse – as well as other women across government, academia, public service, journalism, Māori leadership, and any form of public life. The experiences of prominent Māori individuals show another intersection through which additional violent harm is experienced.

Conclusion

Since the return to Alert Level 4 settings across the country on 17 August 2021, there has been a sharp increase of popularity and intensity of Covid-19 specific disinformation and other forms of 'dangerous speech' and disinformation, related to far-right ideologies. This intensification has included a number of key trends and observations: an increase of both posts and engagement across an ecology of platforms; a shift in reception to the Covid-19 vaccine from vaccine hesitancy to vaccine resistance; the use of memetic material and emotive testimonies; the intensification of language, and the normalisation of that intensification; the use and abuse of Māori motifs and symbols by Pākehā mis- and disinformation spreaders; and the targeting of minority groups and key public figures, particularly those who belong to some intersection of identity.

The ecologies and spread of mis- and disinformation point to a broader threat: that Covid-19 and vaccination are being used as a kind of Trojan Horse for norm-setting and norm-entrenchment of far-right ideologies in Aotearoa New Zealand. Such ideologies include, but are not limited to, ideas about gun control, anti-Māori sentiment, anti-LGBTQIA+, conservative ideals around family and family structure, misogyny, anti-immigration. Mis- and disinformation and 'dangerous speech' pose significant threats to social cohesion, freedom of expression, inclusion, and safety.

References

1. Benesch, Susan, “Vile Crime or Inalienable Right: Defining Incitement to Genocide”, *Virginia Journal of International Law,* (2003).
2. Berentson-Shaw J and Elliot M. *Misinformation and Covid-19: a briefing for media*. Wellington: The Workshop; (2020).
3. Center for Countering Digital Hate. “The Disinformation Dozen” https://www.counterhate.com/disinformationdozen, (2021).
4. Davies, Thom, “Slow violence and toxic geographies: ‘Out of sight’ to whom?”, *Politics and Space,* (2019).
5. 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).
6. Galtung, Johan. “Violence, Peace, and Peace Research”, *Journal of Peace Research* 6, no. 3 (1969): 167-191.
7. Nixon, Rob, *Slow Violence, Gender, and the Environmentalism of the Poor*, (2011).
8. Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism,* New York: NYU Press, (2018).
9. The Dangerous Speech Project*, Dangerous Speech: A Practical Guide*: 19 April 2021 https://dangerousspeech.org/guide/
10. United Nations, *Declaration on the Elimination of Violence against Women,* (1993).
11. Winter, Charlie, Peter Neumann, Alexander Meleagrou-Hitchens, Magnus Ranstorp, Lorenzo Vidino, and Johanna Fürst. "Online Extremism: Research Trends in Internet Activism, Radicalization, and Counter-Strategies." *International Journal of Conflict and Violence* 14, (2020).