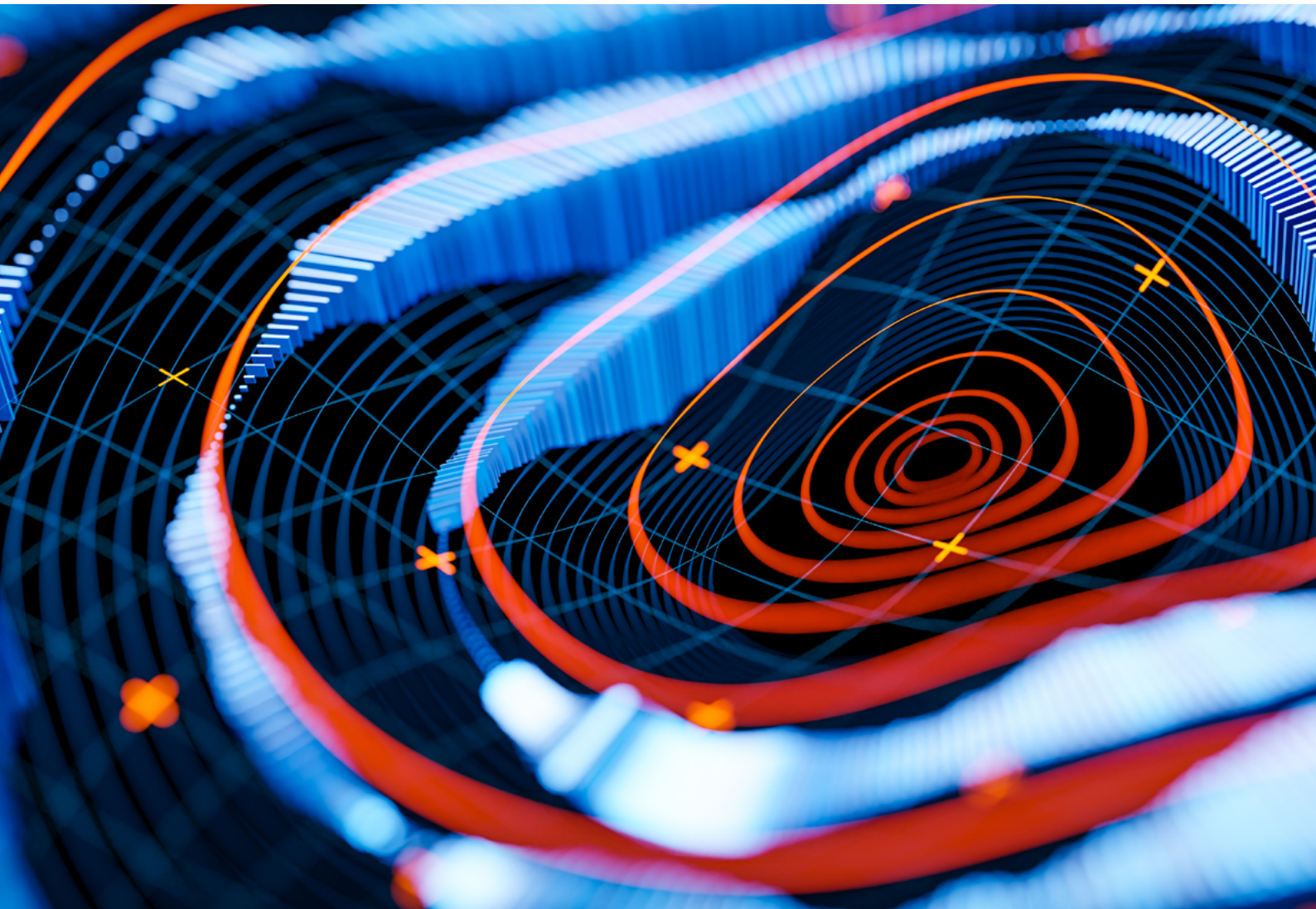
Behavioral Data Science

UTS

Berkeley SCHOOL OF INFORMATION

The Information Integrity Initiative:

Delivering tools and analysis to fight the growing threat of misinformation for Australia



Principal investigator:

Marian-Andrei Rizoiu (UTS) Marian-Andrei.Rizoiu@uts.edu.au
Hany Farid (UC Berkley) hfarid@berkeley.edu
Adam Berry (UTS) adam.berry@uts.edu.au

Key personnel:

Jooyoung Lee Jooyoung.Lee@uts.edu.au
Emily Booth Emily.Booth@uts.edu.au
Elliott Waissbluth ewaissbluth@berkeley.edu

PhD students:

Pio Calderon piogabrielle.b.calderon@student.uts.edu.au
Philipp Schneider philipp.schneider@epfl.ch
Lanqin (Frankie) Yuan lanqin.yuan@student.uts.edu.au

Organisations:

UTS Data Science Institute
University of Technology Sydney, Ultimo, NSW 2007, Australia.

Behavioral Data Science Lab:
https://www.behavioral-ds.science/

School of Information,
University of California, Berkeley, CA 94720, USA.

Period of Performance:

1 July 2022 – 30 June 2023

Layout design by cristiandita.com

Executive summary

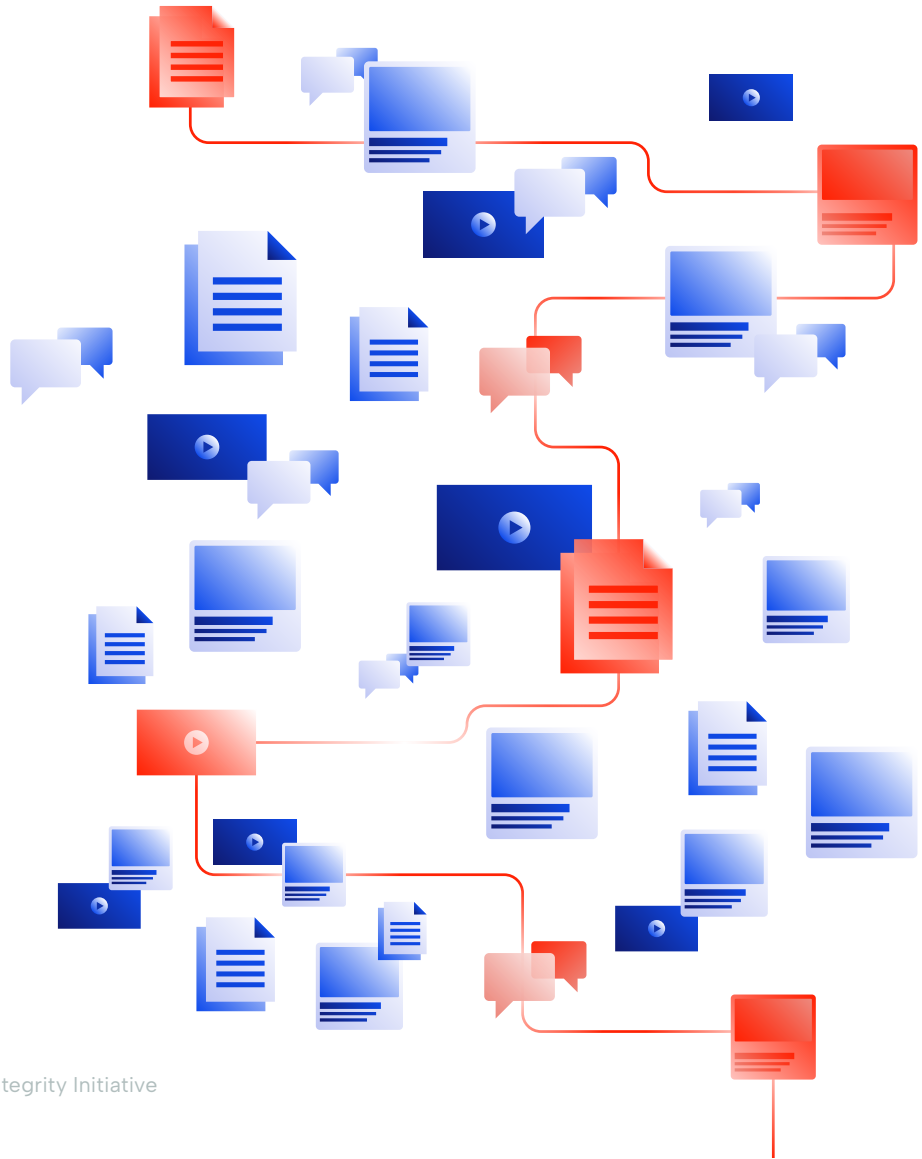
Background and Context

The University of Technology Sydney and the University of California Berkeley are undertaking a research project to investigate the online misinformation landscape in Australia. This three-year project aims to enhance our understanding of the country's sources, transmission methods, and outcomes of misinformation.

The research focuses on retrospective case studies of Australian-specifc online groups, including extreme-leaning and antivaccination ideological groups. The project has three goals:

- Expand knowledge about misinformation flow and consumption in Australia.
- Identify potential responses from public and private entities.
- Develop domestic expertise in dealing with online misinformation while providing training opportunities for Australian talent.

The project is structured into two phases: a 12-month performance period (July 2022 to June 2023), during which the research team works closely with stakeholders, followed by a 24-month research-driven period (July 2023 to June 2025) driven by PhD students exploring fundamental research.



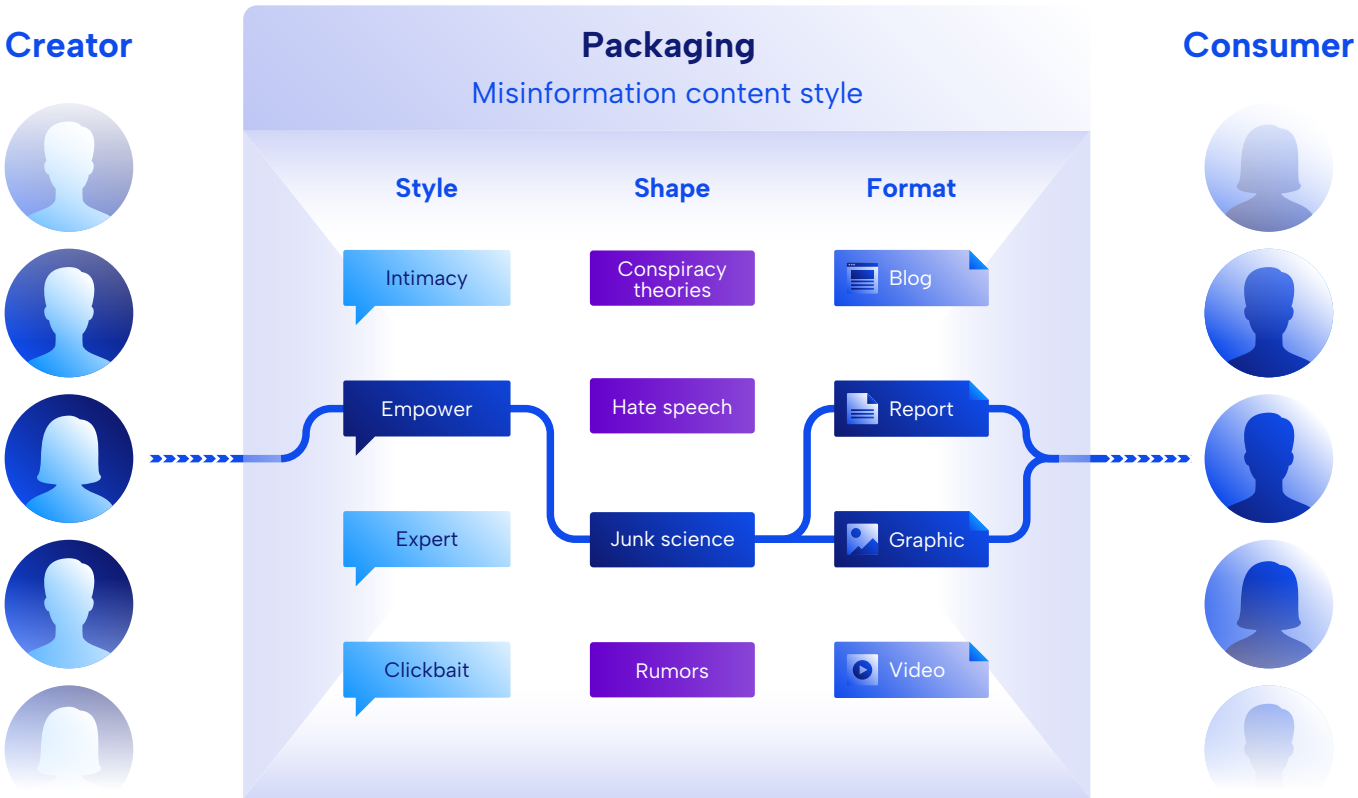
Chapter 1: The Misinformation Eco-System

This research aims to understand how misinformation spreads in Australia's online ecosystem. This research aims to develop an understanding of the misinformation ecosystem in Australia. This expands existing knowledge by taking an integrated approach to producing, funding, spreading and consuming misinformation.

We found that misinformation is created and consumed as an online product targeted to specific vulnerable populations. People tend to believe and share misinformation that aligns with their preexisting beliefs and communication style. Our study suggests that customised messages tailored in the preferred style of the consumers are up to 50% more effective than general communication. Additionally, we can identify different types of users prone to sharing misinformation based on their communication style rather than the content itself.

These findings have important policy implications for countering online misinformation:

1. Create more effective communication campaigns by customising messages for the intended audience.
2. Disrupt the spread of misinformation by defunding advertisements that support it.
3. Detect misinformation producers and consumers by analysing their adopted styles and topics.



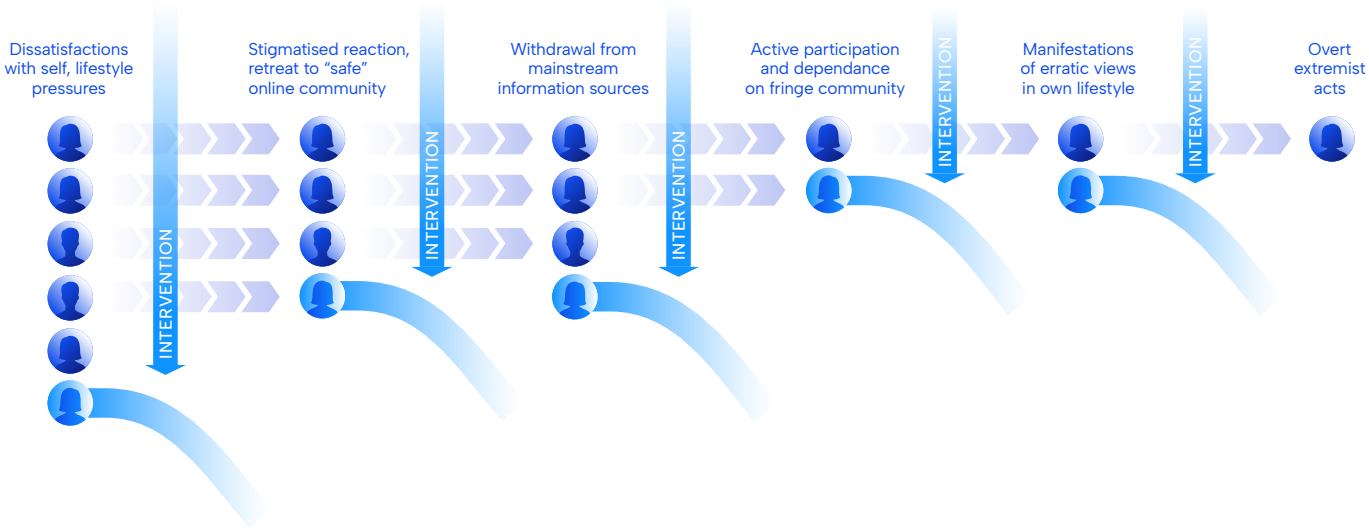
Misinformation is an online product. Consumers have preferences about style, shape and format. Creators cater to these preferences, generating content aligned with the consumer's world views.

Chapter 2: Misinformation Consumption as a "Radicalisation" Pathway

This research aims to understand the process by which online users encounter and adopt misinformation narratives, potentially leading to violent behaviour. We have identified a six-stage pathway of misinformation "radicalisation," ranging from initial confusion and grudges (stage 1) to extreme violence (stage 6). Each stage is associated with specific symptoms, and we have also identified intervention strategies and off-ramps to help individuals disengage from radicalisation.

We find that radicalisation can happen quickly. Not everyone reaches stage 6, and only outliers engage in violent acts. Our model proposes potential off-ramps at each stage, such as compassionate responses from healthcare professionals (stage 2) or understanding partners (stages 3 and 4). We have evaluated our model against existing literature and examined real-life examples of former QAnon sympathisers' journeys within this pathway.

The policy implications of this research include implementing targeted information campaigns and providing specialised training for healthcare professionals to intervene at each stage.



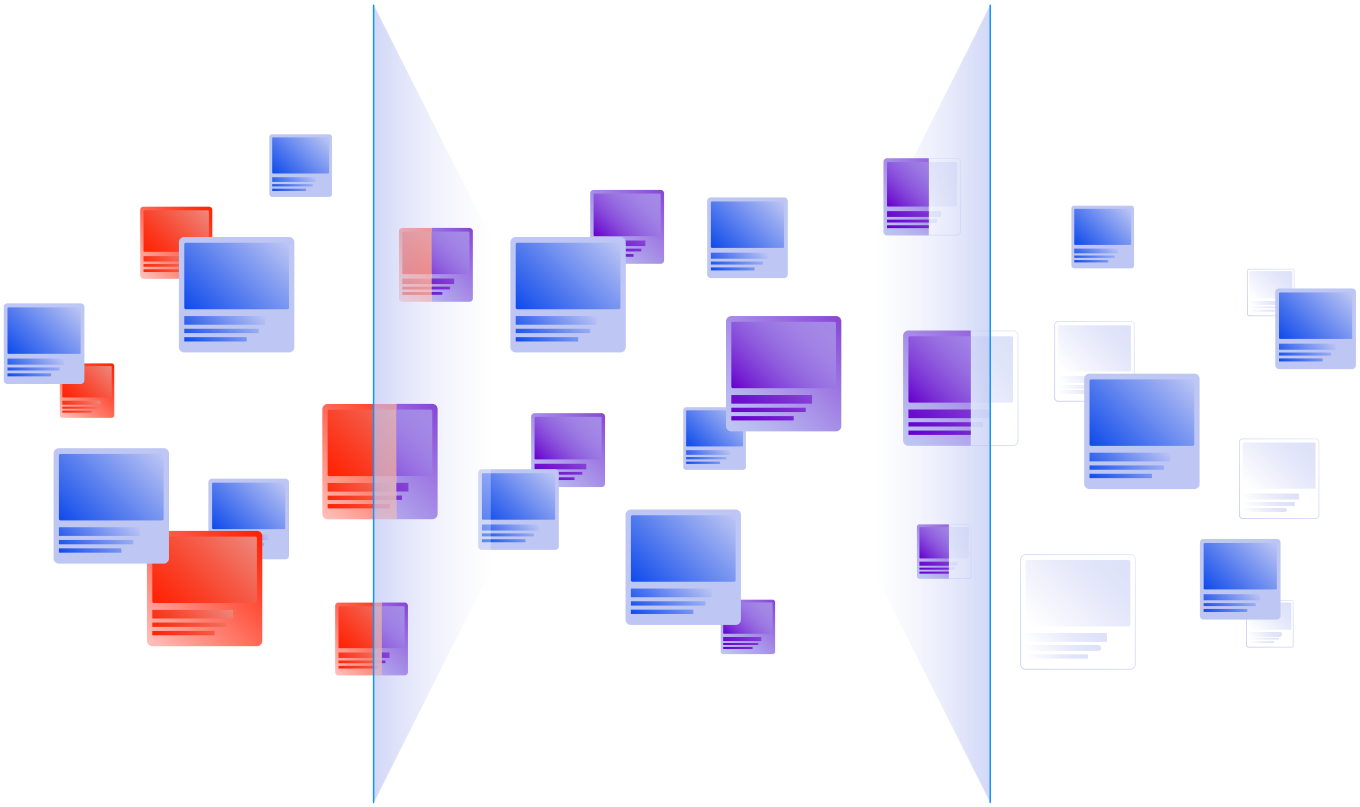
The misinformation "radicalisation" pathway with six stages. Users progress between stages as they get increasingly radicalised. There exist interventions that are effective for each stage. As a result, some users deradicalise shown by the off-ramps.

Chapter 3: The Effectiveness of EU's Digital Services Act

This research examines the effectiveness of the Digital Services Act (DSA). Introduced by the European Union in 2022, the legislation implements notice and action mechanisms (cf. Art. 16) to report illegal online content. It introduces a process for appointing trusted flaggers to detect illegal content (cf. Art. 22). Once such content is flagged, platforms must promptly remove the content. This expands existing knowledge by using state-of-the-art research to analyse the relationship between regulated moderation delay and harm reduction.

The findings demonstrate that even on fast-paced platforms like Twitter, harm reduction can be achieved for highly viral and illegal content. The research provides a method for estimating moderation effectiveness on different platforms. It offers a rule of thumb for selecting content for investigation and flagging, effectively managing flaggers' workload. The results showcase two Twitter case studies; however, the approach applies to other digital service providers.

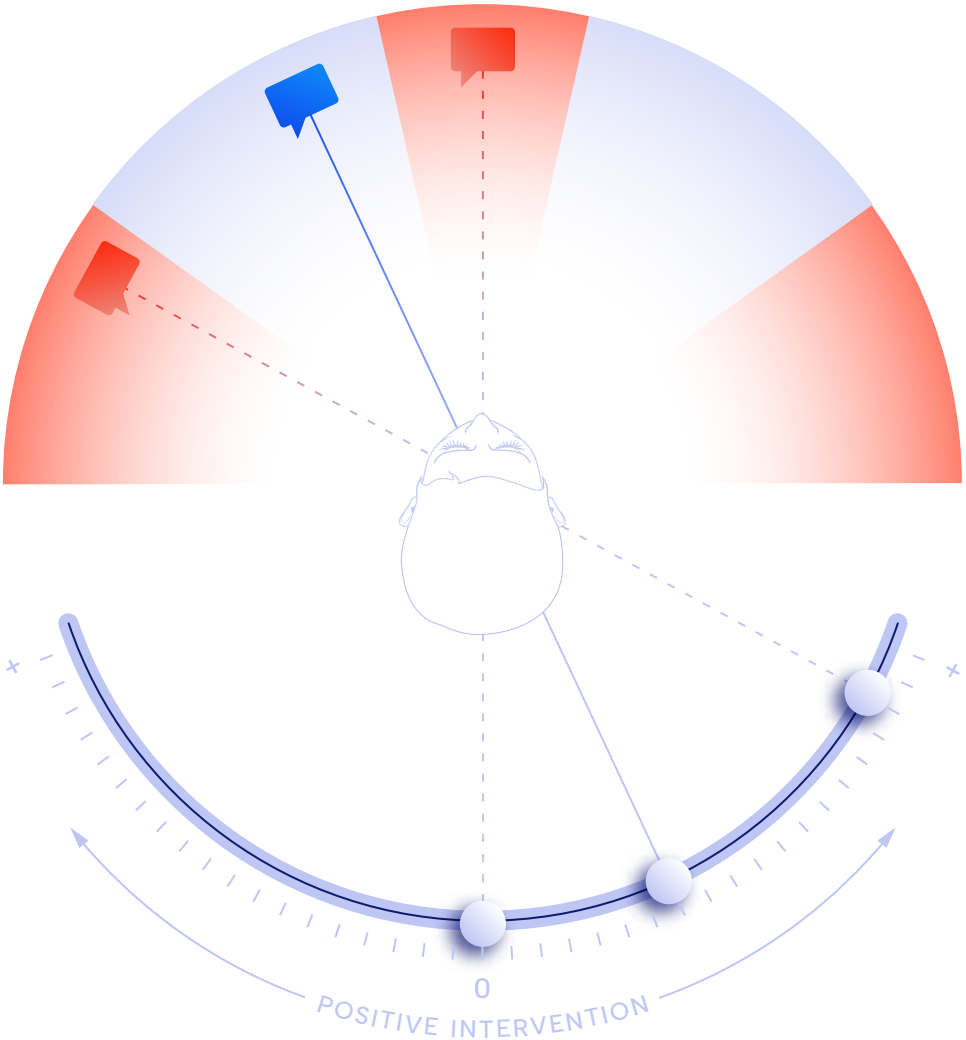
The policy implications of this research suggest that Australia should consider adopting a similar approach to tackle domestic misinformation. Furthermore, the research proposes a prioritisation mechanism for checking and flagging potentially illegal content based on its potential harm, which could reduce workloads. This mechanism is absent in the EU's DSA, making it a potential Australian innovation if implemented.



The DSA moderation works as two thresholds that online posts pass. Illegal content is flagged (left threshold), and the online platforms have 24 hours to remove it (second threshold).

Chapter 4: Synthetic Testbed for Modelling Positive Interventions

This research focuses on the challenging task of designing interventions against misinformation. It explores the concept of positive interventions. Negative interventions involve content moderation and account suspension. However, positive interventions increase attention to debunking opinions 3 that seek to counter misinformation. The research adopts a market model analogy, in which misinformation and debunking opinions compete or cooperate for attention. The findings reveal the interactions between far-right and moderate opinions. For example, debunking far-right ideas on Twitter unexpectedly reinforces instead of countering them. The research offers a way to develop and assess effective interventions to combat misinformation. It produces a testbed to estimate the interventions' impact and the potential to backfire before deploying them in the real world.



Positive interventions orient user attention; however, they need to be carefully calibrated as they can backfire and increase misinformation opinions. Our research builds this calibration.

Chapter 5: Prototype Software to Monitor Online Social Media Discussions and Ads Collection

This chapter presents two software prototypes for monitoring online and social media and ad collection. These tools provide practical approaches to interventions against online misinformation. The first prototype allows monitoring online discussions around specific topics. It features an information dashboard that captures and visualises data from various sources, enabling the definition of topics of interest and the automatic collection of relevant content. The software is extendable and will include automatic labelling of posts using natural language processing techniques and an early detection system for extremist user ideology. The second prototype identifies companies that unintentionally fund misinformation through advertisement. The software extracts ads from the web pages shared by misinformation and extreme-leaning users. The outcomes allow for monitoring online discussions, assessing narratives quickly, and identifying companies that fund harmful content. This provides for intervention against online misinformation by disrupting its funding flow. It has the advantage of helping companies avoid the association between their brand and damaging content.

Our software collects social media and news data on debated topics. It is extensible to allow additional analysis modules, such as detecting extremist ideologies and foreign interference agents and their narratives.

Technical Chapters

1. The Misinformation Eco-System

In this chapter, we define the misinformation ecosystem as consumption pathways and we present our findings around the ecosystem. We further discuss pathways of radicalization in Chapter 2.

1.1 The ecosystem

We define the misinformation ecosystem as a process of misinformation consumption pattern. We show the misinformation ecosystem in Fig. 1.1. Each misinformation consumption pattern connects a consumer to a creator using a distinct packaging of misinformation. In addition, we propose that the proposed misinformation ecosystem treats misinformation as a packaging of style, shape, and format rather than content specific. We further support the proposal through linguistic analysis of misinformational messages in Section 1.3.

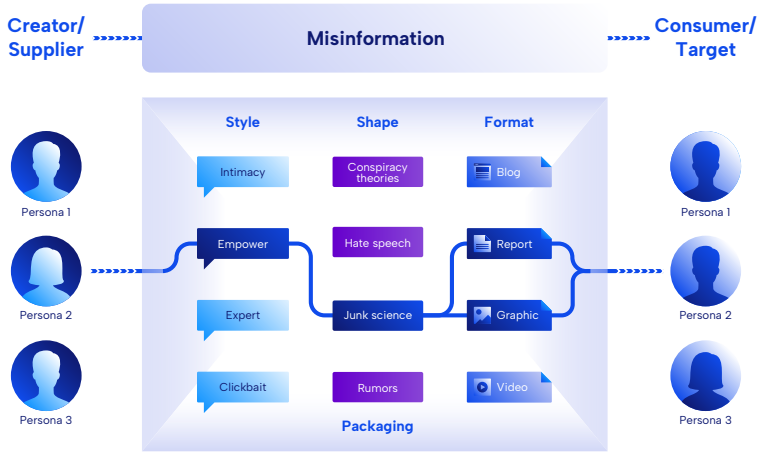


Figure 1.1: The misinformation ecosystem.

When a message containing misinformation is created by a creator, the message is packaged with style, shape, and format to make its pathway to the consumer. For example, Fig. 1.2 shows a misinformation consumption pattern that utilizes a specific misinformation packaging, i.e, “intimacy” as a style of the language, shaped into “rumors”, in the format of “blog” and “graphic” to convey the misinformation.

Figure 1.2: An example pathway of misinformation consumption.

1.1.1 Personae

Building on the work on linguistic features and online observation of misinformation communities discussed in the previous report for this project, we developed three personae to represent the different kinds of people involved in these online spaces. Each persona reflects a different type of online community and concern, embodying the most common attributes (e.g. demographics or lifestyles) observed about commenters in these spaces. The personae were created to humanise and represent these individuals without ever depicting a real person.

The three personae in this project are Jennifer, Patrick, and Aaron. Jennifer is a young mother experiencing fears about childhood vaccination, Patrick is a retired school teacher who is concerned about climate change and the future for young people, and Aaron is a young single man who is interested in supplements as a means to overcome his body insecurities. They correspond with the three misinformation spaces explored in the previous report, of Tay's Way, Info Wars, and Canberra Declaration, supplemented by additional research and observation in different Facebook groups and pages. On the basis of this qualitative observation, we were able to determine the speech style, demographic attributes, and issues of concern for each persona.

The full stories developed for the personae are presented below.

Jennifer

Age:

20–40

Residence:

Australia

Patrick

Age:

65+

Residence:

Suburban NSW

Aaron

Age:

20–35

Residence:

Australia

Jennifer

Age:
20–40

Residence:
Australia

Education:
BA

Occupation:
Formerly a communications specialist for a non-profit, but retired after the birth of her first child

Marital status:
Married with two children (a newborn and a 2 year old)

Jennifer wakes up at 6am each day, often after a broken night's sleep due to her newborn still waking up. She prepares breakfast for her husband and does some yoga, and misses the morning class she used to attend. Jennifer's peace ends when the children wake up, and she spends most of the day juggling their needs. If she can successfully get them to have a nap, she checks Facebook and Instagram to try and keep up with her friend's lives and stay in touch. She might also browse Pinterest, saving pictures to boards about yoga, healthy and quick meals, motherhood advice, and painting tips. Twice a week, she has an afternoon walk with the kids to the nearby supermarket for groceries. In the evening, she makes dinner for her husband. He enthusiastically talks about work, but she feels like she has very little to say about her life and it's not very interesting to him.

Needs

- More sleep
- More certainty on the 'right' things to do for her children
- Regular reassurance about her capabilities to be a mother

Wants

- To spend more spend more time with her friends from work
- To resume a regular 'date night' with her husband when she finds a trusted babysitter
- To pursue her interest in painting with some classes

Values

- Health
- Creativity
- Simple pleasures

Fears

- That she is not being the best mother she could be!
- That her children will get sick
- That her friends at work forget her
- That she's become unattractive to her husband after having two children

Patrick

Age:
65+

Residence:
Suburban NSW

Education:
Bachelor of Education

Occupation:
Retired primary school teacher

Marital status:
Married for 45 years, 3 children, 4 grandchildren

Patrick wakes early and takes a walk around the neighbourhood, before helping his wife up and organising breakfast for her. She's become quite frail, and he has to help her get dressed and walk. After, they often sit outside in the garden and he reads to her. He organises lunch and they both have a short rest, before the grandkids come over after school until their parents finish work. These frantic few hours are their favourite part of the day, even if it gives them a bit of a headache. He used to volunteer at the local library to help struggling children with their reading, but this was closed down due to COVID-19. He helps the kids with homework, and then takes care of chores as his wife supervises them watching TV or spending time in the garden. When the grandkids have been retrieved by their parents he organises dinner, and as his wife watches late night TV, he logs on to the Facebook his daughter signed him up for. He tries to keep up with the posts there, and also finds interesting news items about issues he never hears about on TV. His wife thinks the internet is silly, but he's excited about the new things he can learn there. As he gets ready for bed, he often finds himself still thinking about the things he's read.

Needs

- To see his children and grandchildren often
- To see his GP often for his back pain
- To go to bed early, or he can't think the next day

Wants

- To see his grandchildren grow up to be independent and healthy
- To spend more time outside
- To still be relevant to society

Values

- Humour
- Hard work
- Family

Fears

- Being alone
- Illness and death for him and his wife
- Not understanding new changes in the world
- His family forgetting him
- Financial security due to his dependence on the pension

Aaron

Age:
20–35

Residence:
Australia

Education:
Law student (final year)

Occupation:
Summer law clerk

Marital status:
Single

Aaron wakes at around 8am and makes a mad rush for either university or his summer job, skipping breakfast. He is enthusiastic about both and works hard at each. By lunch time, he is feeling exhausted and deserving of a "treat", often buying a burger or pizza meal–and after all, he skipped breakfast, which makes up for the extra calories. He works intently through the afternoon, and leaves work feeling exhausted again. He often tells himself that he's too tired to hit the gym now, but he'll go tomorrow. On Fridays, he spends the night with his Church youth group, but it's increasingly bittersweet, as he will have to leave it after his upcoming birthday. On nights when he doesn't have Church, he often winds up having a few beers and scrolling through Reddit, where he can vent his frustration about still being single and find other interesting blogs to explore.

Criteria for success

He feels most successful when he imagines himself working as a real lawyer to support a wife and kids

Needs

- Someone to talk to about his anxiety over his scrawny appearance, compared to action heroes in Hollywood
- His Church youth group, which he's getting too old to be a member of

Wants

- A girlfriend
- To be good at surfing, even though he's embarrassed by how bad he is and sunburns easily
- To get a job when he finishes university

Values

- Strength and confidence in men; beauty in women Christianity
- Recognition for his hard work in his studies and at his holiday job

Fears

- Professional failure: He will fail as a lawyer, disappointing his parents
- Personal failure: He won't get a girlfriend
- He won't ever become a father
- He will lose his connection to his Church after "graduating" the youth group he cherishes

1.2 Curated datasets

In this section, we present datasets from three online platforms and show how they can be intersected for comparison.

1.2.1 News publishers

We collected news articles that are indexed by Google news through The Daily Edit application. The collection period for the news articles range from November 9, 2022 till January 11, 2023. There are 9929 news publishers in this dataset.

Table 1.1: In this table, we report the number of extracted urls from the tweets and Facebook posts as well as the proportion of urls from the identified news publishers in Section 1.2.1. Additionally, we show top 10 most shared identified news publishers for each group.

	Twitter	Facebook	
		Far right	Antivax
#Total urls	**1827162**	**846**	**556**
#Identified	**3377 (0.2%)**	**62 (7.3%)**	**20 (3.6%)**
	Top 10 identified publishers		
1	www.theaustralian.com.au	www.skynews.com.au	spectator.com.au
2	www.smh.com.au	www.couriermail.com.au	www.smartcompany.com.au
3	www.abc.net.au	www.dailytelegraph.com.au	www.ncbi.nlm.nih.gov
4	www.heraldsun.com.au	www.news.com.au	www.theepochtimes.com
5	theguardian.com	www.abc.net.au	www.smh.com.au
6	nypost.com	www.smh.com.au	www.news.com.au
7	www.dailymail.co.uk	7news.com.au	hellocare.com.au
8	www.theage.com.au	www.portnews.com.au	www.9news.com.au
9	www.afr.com	www.stuff.co.nz	www.abc.net.au
10	www.news.com.au	www.dailymail.co.uk	humanrights.gov.au

1.2.2 Twitter users

We have curated list of far right Twitter users collected from previous research. We collected 3200 most recent tweets (as allowed by Twitter API) from the 1496 identified far right Twitter users. 1, 827, 162 urls were shared by the users in the tweets and 3377 news publishers from Section 1.2.1 appeared in the urls.

1.2.3 Facebook groups

We assembled two lists of Facebook groups for specific ideologies, namely Australian Antivax and Australian Far right groups. We collect posts from the Facebook groups in the lists via Crowdtangle API. We retrieve

Facebook postings within the identified groups from January 23, 2021 to February 24, 2023. 6017 posts were collected from the Far right groups and 2969 posts were collected from the Antivax groups. Also, 846 urls extracted from the posts by Far right groups and 558 urls extracted from the posts by Antivax groups. In Table 1.1 we summarize the intersection of the datasets introduced in Section 1.2.

1.3 Linguistic features as identifiers of extreme groups

Previous report focused on ideological differences in association to the media bias, sharing of misleading texts, and linguistic features from Linguistic Inquiry and Word Count (LIWC). In this section, based on our proposed ecosystem of misinformation Section 1.1, we focus on capturing the packaging of misinformation rather than the content of misinformation. We hypothesize that misinformation is most effectively delivered to target audience via a packaging shown in Fig. 1.1, not by means of contents, e.g., anti-vaccination.

Table 1.2: Summary of the linguistic tools used in this section.

	LIWC	Grievance	StyloMetrix
Summary	LIWC is a transparent text analysis program that counts words in psychologically meaningful categories	Grievance dictionary assess grievance-fuelled communications through language	StyloMetrix is a tool for creating text representations as StyloMetrix vectors
#features	**89**	**22**	**175**

1.3.1 Linguistic measurements: LIWC, GRIEVANCE, STYLOMETRIX

In this section, we introduce three linguistic metrics which quantifies linguistic features in text. We chose these three metrics to present because 1) LIWC is one of the most widely used text analysis tool in psychology and recently adopted by computational social scientists to draw insights into human behavior through more computational methods [1], 2) a study about militant right-wing extremism [2] suggests that a critical ingredient of militant extremist mindset that distinguishes it from social conservatism is grudge which can be captured by GRIEVANCE dictionary, and 3) STYLOMETRIX, a grammar-related statistical representation of text, shows that STYLOMETRIX without the semantic layer is sufficient to detect the genre of the text [3].

LIWC

LIWC dictionary includes both content (e.g., death, religion) and function (e.g., conjunctions, articles) words. LIWC (version 2022) has 117 features in total but in order to capture stylistic differences of the extreme groups, we removed content related features such as religion and family.

GRIEVANCE

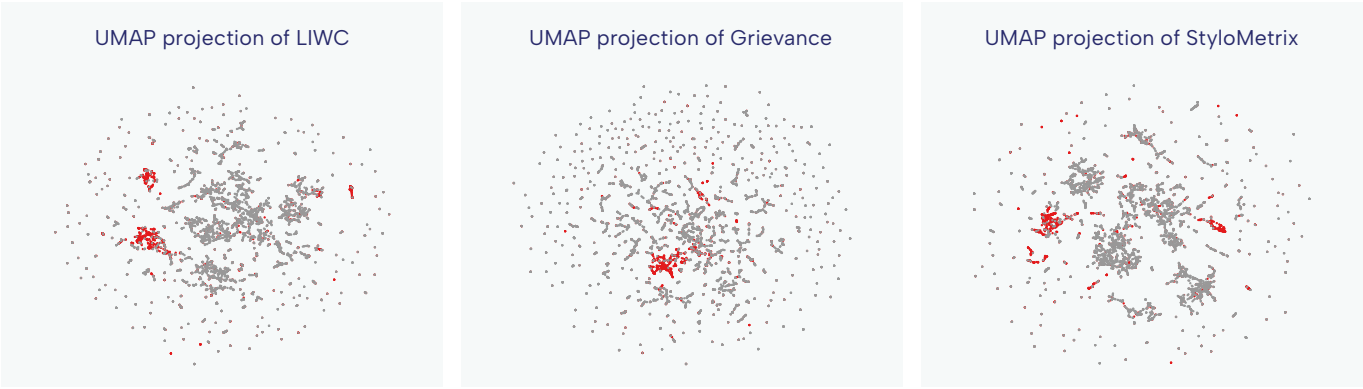
GRIEVANCE is a psycholinguistic dictionary that can be used to automatically understand language use in the context of grievance-fueled violence threat assessment [4]. We include GRIEVANCE dictionary even though it is content based measurement due to its ability to extract violence and threat specific words.

STYLOMETRIX

STYLOMETRIX is a tool that allows for representing a text sample of any length with a linguistic vector of a fixed size [3]. It has several preferable characteristics over other well known contextual embedding such as BERT. First, STYLOMETRIX vectors encode entire documents resolving the issue of varying text lengths. This could help when combining texts from multiple platforms, such as Facebook and Twitter. Second, STYLOMETRIX vectors encode the stylistic structure of the entire sample, not the "meanings" of the words. In [3], STYLOMETRIX showed a good performance on content classification using the style vectors, STYLOMETRIX.

In Fig. 1.3, we compare two dimensional representation of the three linguistic measures (LIWC, GRIEVANCE and STYLOMETRIX) using UMAP. The original dimensions of each measure is in Table 1.2. Since there are 6017 posts from **Far right** and 2969 posts from **Antivax**, there are about twice as many gray dots as red dots in each figure. In all three measures, we observe that posts from Far right group are more scattered while posts from **Antivax** group are more congregated. Based on our misinformation ecosystem, this implies that there exist more diverse packaging of misinformation in **Far right** posts than **Antivax**. This also corresponds to the nature of these two extreme groups in that **Antivax** group is focused on a specific topic of "vaccine" whereas **Far right** group can talk about various topics.

Figure 1.3: UMAP projection of Facebook posts from Far right and Antivax groups using embedding of LIWC, GRIEVANCE and STYLOMETRIX respectively. Gray color represents posts from Far right and red color represents posts from Antivax.



1.3.2 STYLES as fingerprints

Style words reflect how people are communicating, whereas content words convey what they are saying. It is suggested that style words are much more closely linked to measures of people's social and psychological worlds [5]. Styles encompass a range of linguistic features, including sentence structure, grammar, and punctuation patterns. Unlike the content of the text, which can be influenced by subject matter, external

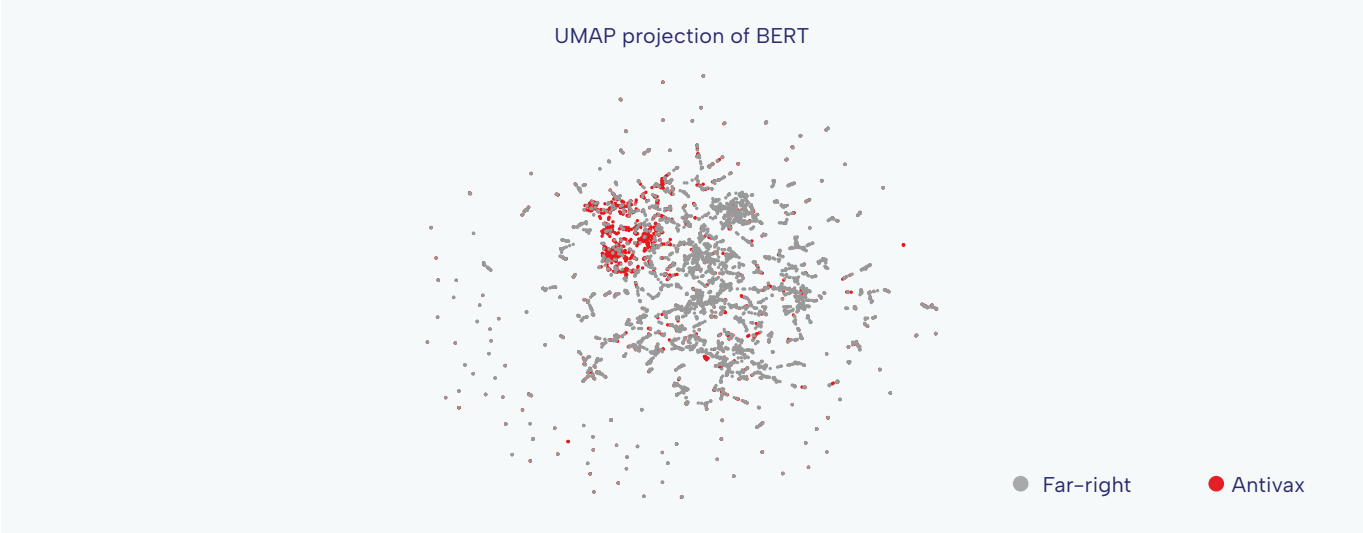
sources, or intentional deception, stylistic features are intrinsic to one's writing style and are less prone to deliberate manipulation.

In this section, we focus on stylistic differences because the contents or the topics of interest within a group can dynamically change and but stylistic fingerprints persist [6]. Another reason to consider stylistic features over content related features is that stylistic representation (or embedding) is scalable while content related features are constrained to specific topics. In the following subsections, we show classification performance on distinguishing the extreme groups first and then distinguishing the different styles of texts regardless of the groups.

STYLE based classification of groups

We design an experiment to compare the performance of the three dictionary measures (LIWC, GRIEVANCE and STYLOMETRIX) in distinguishing posts from the extreme groups. In addition, we add another popular text encoding technique, BERT [7]. BERT, Bidirectional Encoder Representations from Transformers is a machine learning technique for natural language processing. It is used as a baseline in many research experiments due to its high performance. As opposed to the other three dictionary based encoding, BERT takes into account the context for each occurrence of a given word. BERT encodes a text from the left and right side during training (thus bidirectional), and consequently gains a deep understanding of the context.

Figure 1.4: Umap projection of Facebook posts from **Far right** and **Antivax** groups using BERT embedding.



In Fig. 1.4, we show the two dimensional projection of BERT embedding for Far right and Antivax posts. It is interesting to see that BERT and GRIEVANCE embedding (in Fig. 1.3) shows Antivax posts (red dots) as a one big group while LIWC and STYLOMETRIX embedding shows Antivax posts as multiple clusters. This also suggests that STYLE based encodings can detect different STYLES within the same group which is useful to monitor radicalization pathways and interventions.

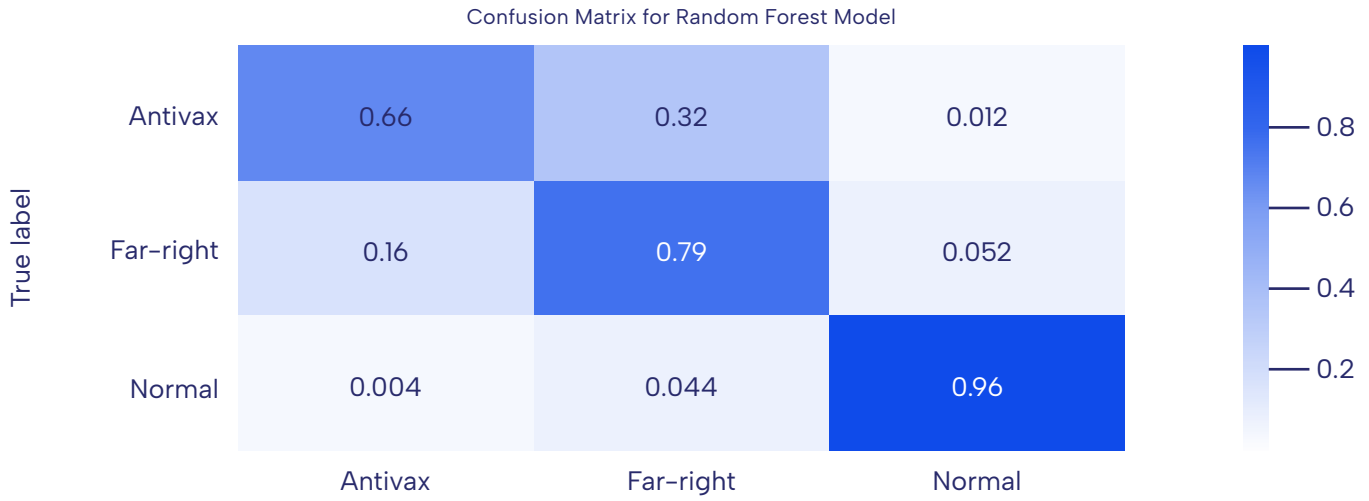
Table 1.3 reports classification performance using three basic classifiers. The reported performance is the average result of 10-fold cross validation. Since BERT is pre-trained to learn precise context of the given text, it performs the best most of the time. LIWC, GRIEVANCE and STYLOMETRIX are dictionary based measures

which calculates percentage of the features used in the given text. Since contexts can change overtime, we want to rely on STYLE of text which is more robust. On the most right column, we add another simple measure, LGS, which is a concatenation of LIWC, GRIEVANCE and STYLOMETRIX. Using Random Forest classifier, LGS outperforms BERT. This preliminary results suggest that with a larger training dataset and finetuning, STYLE based measures can perform better than BERT in distinguishing texts from different groups. In other words, without learning the context of a given text, we can predict which group the post came from based on STYLE of the text.

Table 1.3: Performance of three group classification. In this setting, on top of the two extreme groups, **Far right** and **Antivax**, we also add normal Facebook group. As a result, there are 3 groups to classify into, each group has 1000 randomly sampled posts.

		BERT	LIWC	Grievance	StyloMetrix	LGS
Logistic regression	accuracy	0.80	0.73	0.56	0.69	0.74
	macro F1	0.79	0.73	0.54	0.68	0.74
Linear SVC	accuracy	0.76	0.66	0.57	0.71	0.68
	macro F1	0.76	0.61	0.55	0.70	0.64
Random Forest	accuracy	0.75	0.72	0.65	0.70	**0.77**
	macro F1	0.75	0.71	0.63	0.69	**0.76**

Figure 1.5: Confusion matrix for the three group classification using Random Forest on LGS features.



In Fig. 1.5, we present the confusion matrix for the three group classification using Random Forest classifier with LGS vectors which outperformed BERT in Table 1.3. Based on the confusion matrix, normal Facebook posts correctly classified 95% of times whereas Far right and Antivax are only correctly classified 79% and 66% of times respectively. This implies that we can easily separate normal group from the extreme groups using the STYLE of the text.

STYLE based classification of packaging

In this section, we attempt to classify STYLES of texts regardless of the groups. This preliminary experiment uses manually annotated text samples from Far right and Antivax groups. We annotated STYLES of 62 sample posts. The identified STYLES in these samples are "Casual", "Empowerment", "Clickbait", "Expert" and "Intimacy". This annotation is a result of sampling 50 posts from both Far right and Antivax then removing the posts which did not have any STYLES. The number of posts for each STYLE is shown in Table 1.4.

Table 1.4: Number of posts per STYLE.

	Casual	Empowerment	Clickbait	Expert	Intimacy	Total
Far-right	10	23	0	12	1	46
Antivax	11	4	1	0	0	16
	21	27	1	12	1	62

Table 1.5: Classification result of "Casual" style from others.

"Casual"		BERT	LIWC	Grievance	StyloMetrix
Logistic regression	accuracy	0.74	0.59	0.66	0.64
	macro F1	0.68	0.50	0.40	0.46
Linear SVC	accuracy	0.72	0.56	0.66	0.56
	macro F1	0.68	0.48	0.40	0.44
Random Forest	accuracy	0.66	0.54	0.56	0.61
	macro F1	0.50	0.37	0.36	0.44

Table 1.5 to Table 1.7 show classification performance of "Casual", "Empowerment" and "Expert" STYLE respectively. "Clickbait" and "Intimacy" STYLES have only 1 posts for each thus cannot apply classification. As shown in Table 1.5, BERT outperforms for all the other metrics for all three classifiers to classify "Casual" posts from others. But for "Empowerment" and "Expert", BERT did not outperform in all cases and the even for the cases where BERT outperforms, the difference between BERT and the second best performing metric is smaller. This suggests that "Casual" is more context aware STYLE compared to "Empowerment" and "Expert". Among the three STYLES that we examined, "Expert" achieves the highest accuracy in all cases. This could be due to the fact that "Expert" posts tend to follow certain formats such as adding references and citing other sources. On the other hand, "Empowerment" STYLE is shown to be the hardest to classify (Table 1.6). "Empowerment" can be conveyed in many different ways including direct and explicit statements, while others may rely on subtler cues, metaphors, or storytelling. This variability in expression makes it challenging to define a precise set of features that consistently characterize empowerment style text. We believe that by adding SHAPE and FORMAT to the classification will improve the performance.

Table 1.6: Classification result of "Empowerment" style from others.

"Empowerment"		BERT	LIWC	Grievance	StyloMetrix
Logistic regression	accuracy	0.54	0.48	0.59	0.43
	macro F1	0.51	0.45	0.37	0.38
Linear SVC	accuracy	0.51	0.49	0.59	0.51
	macro F1	0.48	0.46	0.37	0.49
Random Forest	accuracy	0.56	0.44	0.44	0.53
	macro F1	0.52	0.41	0.40	0.50

Table 1.7: Classification result of "Expert" style from others.

"Expert"		BERT	LIWC	Grievance	StyloMetrix
Logistic regression	accuracy	0.80	0.74	0.79	0.79
	macro F1	0.66	0.67	0.44	0.44
Linear SVC	accuracy	0.82	0.74	0.79	0.82
	macro F1	0.72	0.67	0.44	0.58
Random Forest	accuracy	0.79	0.80	0.75	0.80
	macro F1	0.44	0.52	0.51	0.52

1.3.3 Real-world experiment: effectiveness of STYLIZED advertisements

In this section, we test our hypothesis that misinformation is a packaging (style + shape + format) rather than a content. Specifically, in this preliminary setting, we first test the effect of utilizing STYLED text in publicising information.

Experiment design

We developed a series of Facebook advertisements targeted towards our personae to determine the impact of language style on these different communities in relation to issues of concern. Facebook has an A/B Testing program, which allows for two different versions of the same ad to be tested with the same demographic to determine what achieves the best result. Our metrics for success were the clickthrough rate and cost per click. We designed a series of ads targeting our personae about an issue we determined to be relevant to them, as well as a neutral, unstyled ad designed to resemble the tone and structure of a government ad. These unstyled ads used simple declarative statements that were unemotional and more formal in tone. Copyright free images were sourced to accompany each ad, and were identical for both the styled and unstyled versions

of each ad, eliminating visuals as a factor of influence. An Australian government website was used as the landing page for any clicks on the ads, to ensure individuals landed on a website with reliable information about the topic of their interest. The ads were associated with a Facebook page connected to the Behavioural Data Science Lab at UTS.

We ran these A/B tests for Jennifer, Patrick, and Aaron for one week each. Jennifer's ad used the "Intimacy" STYLE shown in Fig. 1.1, with the topic of childhood vaccination. We determined that this would appeal to Jennifer as it would indicate empathy and closeness for her circumstances, and that a barrier to clicking on the neutral unstyled ad could be the more formal, masculine nature of the language. Patrick's ad used a more authoritative "Expert" STYLE approach designed to encourage the reader to feel they should be in control of knowing "the facts". Patrick's topic was climate change. It was hypothesized that his barrier to clicking on an unstyled ad might be a belief in his own knowledge and life experience. For Aaron, the style used for his ads was that of "Empowerment", with his text inviting him to become a better version of himself. It was thought that his barrier to the unstyled ad may be the authoritative tone, against which a young man like Aaron might chafe. His topic was supplements, as an asset for physical fitness. Detailed descriptions for these personas are in Section 1.1.1.

Table 1.8: Summary of Facebook Ad outcome report. Impressions is the number of views on the ad. CTR is Click-Through-Rate (ratio clicks to impressions). Engagement is the total number of actions people take on the ad (shares, reactions, saves, comments etc.). The winner ad is shown in bold font (no winner on CTR% for Aaron).

			Facebook Ad			
Persona	**Style**	**Topic**	**Impressions**	**Clicks**	**CTR%**	**Engagement**
Jennifer	Intimacy	childhood	13107	129	**0.98**	**135**
	Unstyled	vaccination	15567	101	0.65	112
Patrick	Expert	climate	10082	334	**3.31**	**468**
	Unstyled	change	10095	170	1.68	370
Aaron	Empowerment	supplements	17706	118	0.67	**121**
	Unstyled		14552	102	0.71	104

Experiment outcomes

We show the summary of Facebook ad results in Table 1.8. Each Persona had two campaigns, i.e., STYLED and unstyled version. Impressions is the number of times that ads were on screen. Impressions is a common metric used by the online marketing industry. Impressions measure how often the ads were on screen for the target audience. Clicks is the number of clicks including link clicks as well as clicks on other parts of the ad. CTR (Click-Through Rate) is percentage of times people saw the ad and performed a link click. The metric is calculated as clicks divided by impressions. Post engagements are the total number of actions that people take involving the ads on Facebook. Post engagement includes all actions that people take involving the ads while they're running. Post engagements can include actions such as reacting to, commenting on or sharing the ad, claiming an offer, viewing a photo or video, or clicking on a link.

For Jennifer and Patrick, STYLED version campaign had higher CTR compared to the unstyled versions. Unstyled campaign for Aaron resulted higher CTR than STYLED version. However, for all three, STYLED version of ads attracted more engagement. The fact that only “Empowerment” ad showed lower CTR maybe due to more subjectivity in “Empowerment” compared to other STYLES due to its motivational nature. Note that “Empowerment” was the most difficult to distinguish from others using classifiers (Table 1.6).

1.4 Summary and Discussion

In this chapter, we proposed the misinformation ecosystem as a pipeline. We argue that misinformation is a packaging which wraps information or content to target consumers. In order to explore how the packaging is strategically targeting consumers, we created representative personae. We use them in Facebook ad experiment to test effectiveness of packaged (STYLED) contents. In order to identify the packaging of information, we focused on detecting STYLES of texts. We showed that, with only stylistic features, we can distinguish texts from the extreme groups (Far right and Antivax) as good as BERT, the state-of-the-art context aware encoding. We plan to investigate SHAPE and FORMAT, which are the other two factors of the packaging, along with STYLE to improve the classification performance.

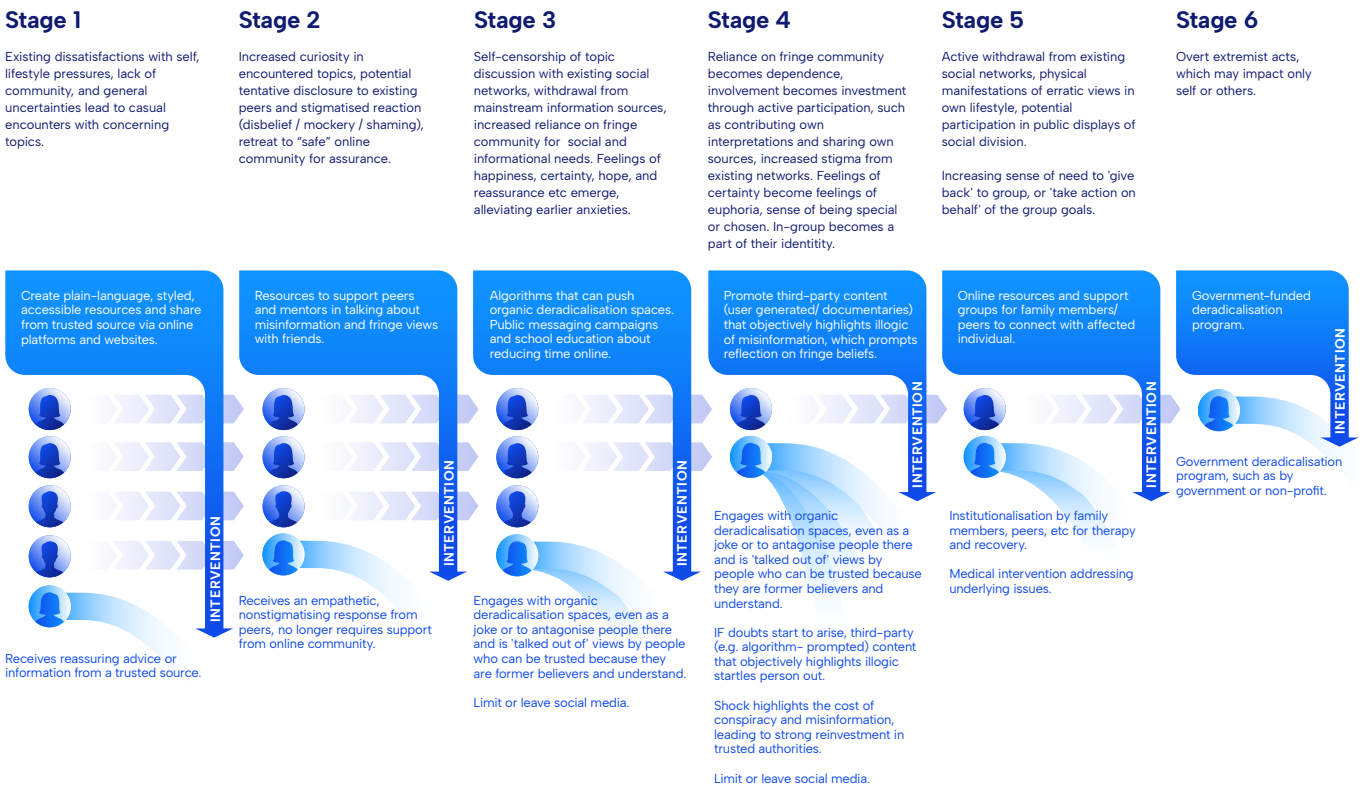
2. Misinformation Consumption as a "Radicalisation" Pathway

2.1 Introduction: Radicalization pathways overview

Our proposed model represents a staged process by which people become radicalised that is relevant to contemporary threats. Many existing models or pathways are highly tailored to their research focus, which was observed in our literature search stage to be either about Jihadist radicalisation or "lone wolf" terrorists. While useful, the focus on context-specific behaviours that signal such radicalisation means that other forms of radicalisation cannot be captured in these models, such as the adoption of extreme fringe views related to vaccines. Our proposed pathway aims to be more general in nature, tracing the behaviours, mental state, and types of triggers that can induce and advance radicalisation. It is intentionally general in nature when describing these traits, with the goal of being broad enough to encompass a range of "topics" about which one may adopt fringe views. We suggest that this pathway could be used to understand and humanise radicalised individuals, to evaluate how far they have been drawn into fringe views, and determine what interventions could be appropriate.

The pathway in a nutshell. The model charts the transition from existing anxieties and a kind of catalyst that sees the individual seeking out some form of stability through information, and the resulting stages they may progress through as misinformation or conspiracy theories act as a balm for their stressors. Fig. 2.1 summarises the stages of the pathway.

Figure 2.1: Summary of the radicalisation pathway with six stages.



In stage one, an individual with uncertainties encounters misinformation of some kind that triggers them to entertain these ideas. In stage two, the individual attempts to discuss their tentative new misinformation-influenced views with a peer or mentor, and experiences a stigmatising response. In stage three, the individual spends more time learning about their new interest. In stage four, the individual is immersed in both the misinformation and the related community. In stage five, an individual may exhibit public erratic behaviour in support of their views and against contrasting views. In stage six, the final stage, the individual undertakes an extreme act that causes harm to either oneself or another person. An individual may not necessarily reach the later stages; nevertheless, they may still experience negative impacts on their own lives, or cause their own loved ones considerable concern. We fully detail each of the stages in Section 2.3.

2.2 Methodology and data sources

This pathway model was developed based on existing scholarship, as well as research looking at former members of far-right, QAnon, and anti-vaxxer communities, who for the most part, departed from their radicalised communities before reaching the extreme final stage. The search of existing scholarship was completed on 28 February 2023, using the University of Technology library catalogue. Rather than conducting a systematic literature review to survey the related literature as a whole, we aimed to find existing models or pathways for radicalisation which could enrich and support the pathway we had developed on the basis of our research. There were two primary searches conducted, which were " 'radicalisation path*' OR 'radicalization path*' or 'radicalisation journey' " (32,322 results) and " 'radicalisation model' OR 'radicalization model' " (110 results). Journal articles, book chapters, and dissertations were included, while other non-scholarly formats such as newspaper articles and reviews were excluded from results. The language was limited to English. For the former search, due to its large volume of search results, we reviewed 100 items, until the relevance of results veered away from the search focus. For the second search, all 110 results were reviewed for potential inclusion. Ultimately, from these two searches, 55 results were downloaded for close reading. During this stage, articles were read and considered for potential inclusion on the basis of the following criteria:

- The articles either outlined some kind of staged process of radicalisation, or closely examined a potential contributing factor in-depth
- The article was not about a purely computational or simulated model
- The model proposed held relevance for our own model; that is, its insights were applicable to the forms of radicalisation that we examined
- The models themselves had been developed by scholars, and not any government or other organisations which could be tailored toward their own context and needs

Papers were tabulated in an Excel spreadsheet upon determining that they were relevant. Of all the results surveyed, there were nine papers that were ultimately included. Within this were six which outlined a staged process for radicalisation that we identified as comparable to our own, and an additional three papers that provided detailed insight into relevant forces that could enable the progression of radicalisation.

The active research for this project is drawn from two years of online observation of fringe groups by Emily Booth under a past project of CI Rizoiu, further developed in this project. During observation periods, Facebook Groups and Pages manually identified as sites where misinformation was consistently posted were reviewed on an ongoing basis, with comprehensive field notes taken on post content, style, and motivations

for involvement in the misinformation in the community. This initial observation process formed the basis of the first draft of the pathway. This was further enriched by a second wave of research. This included close study of 27 news reports in which former believers of misinformation and conspiracy theories spoke about their journeys. These were coded using NVivo software to identify behaviours and views in accordance with stages one-to-six. Additionally, we sourced 20 posts from online, organic deradicalisation spaces on social media platforms, where former and doubting believers posted about their journeys and efforts to leave their fringe communities. These posts were de-identified and then also coded in NVivo in accordance with stages one-to-six. These three layers of original research were then mapped against existing models found in past scholarship to develop the model to its current state.

2.3 Our Proposed Model: the Pathway

2.3.1 Stage 1

The preconditions that make individuals vulnerable to radicalisation are diverse and numerous, but have been found to have some underlying commonalities in previous studies. These can include an individual having some kind of dissatisfaction with themselves, experiencing lifestyle pressures, having a lack of community, and/or general uncertainties about life. Kruglanski et al. [8] identify three main influences that lead individuals to be vulnerable to radicalisation and conspiracy theories. These include a need for significance and a sense of importance to the world, a need for "narrative", or a worldview that gives structure, meaning, and purpose to one's existence; and a need for a network and sense of community [8]. Similarly, Klausen et al. [9] argue that a pre-radicalisation stage is marked by a combination of disillusionment with the world, a personal crisis such as a death in the family or drug addiction, and the start of an information-seeking process with existing or new authority figures. This combination of a recognition of some kind of personal crisis with some kind of political dis/engagement is routinely noted in the literature. Vergani et al. [16] label these as "push factors"—political, structural, or sociological forces—and "pull factors"—personal factors and individual experience—which work in tandem to make someone vulnerable to radicalisation. When one or both are inflamed, such as by an incident like the "transformative triggers" described by Winter and Feixas [11] that undermine an individual's self-concept and worldview, a person is then a prime target for radicalisation.

Similarly, Catherine and Louis [14] and van Eerten et al. [17] both recognise personal vulnerabilities [14] or sensitivities [17] as prime pre-conditions for radicalisation, which may co-occur with some minor exposures to potentially radicalising content. Pepys et al. [15] identify these same factors, but draws attention also to the necessity of an individual's "exposure to radicalising moral contexts, and the emergence of radicalising settings" to actually trigger the radicalisation process. These settings are often political or religious in nature, and can include a newly discovered context or a change in one's existing networks. McCauley and Moskalenko [12] consider personal victimization and political grievance to both be powerful forces that can make an individual vulnerable, while El-Muhammady [10] pairs features like social isolation and low-self worth with more personal motivations to action such as experiencing cognitive distortions, seeking quick or easy means of creating change for the better, and high levels of narcissism. It is important to note therefore that while existing research has observed these character traits and experiences as factors as preconditions to radicalisation, they are not in and of themselves signs of fringe behaviour or views; but rather, potential traits that may increase susceptibility.

Table 2.1: Mapping existing literature on our proposed radicalisation pathway

	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5	Stage 6
Paths and models from scholarly literature	Existing dissatisfactions with self, lifestyle pressures, lack of community, and general uncertainties lead to casual encounters with concerning topics	Increased curiosity in encountered topics, potential tentative disclosure to existing peers and stigmatised reaction (disbelief/ mockery/shaming), retreat to "safe" online community for assurance	Self-censorship of topic discussion with existing social networks, withdrawal from mainstream information sources, increased reliance on fringe community for social and informational needs	Reliance on fringe community becomes dependence, involvement becomes investment through active participation, such as contributing own interpretations and sharing own sources, increased stigma from existing networks	Active withdrawal from existing social networks, physical manifestations of erratic views in own lifestyle, potential participation in public displays of social division	Overt extremist acts, which may impact only self or others
Terrorism and conspiracy theories: A view from the 3N model of radicalization [8]	Need for Significance, Narrative, and Network					
Toward a Behavioral Model of "Homegrown" Radicalization Trajectories [9]	Pre-Radicalization			Stage 1: Detachment	Stage 2: Peer-Immersion and Training	Stage 3: Planning and Execution of Violent Action
Malaysian radicalisation model [10]	S1. Pre-radicalisation life: period before their exposure to radical ideas, individuals and experience. AND S2. The first contact with radical ideas, individuals and experience		S3. The process of learning and deepening or radical ideas, experience and events		S4. The process of externalising extremist narratives. The subject tends to share their knowledge, ideas, experience	S5. Actualisation phase refers to participation in the actual act of terrorism, not just verbal expression or promotion
Toward a Constructivist Model of Radicalization and Deradicalization [11]	1. The radicalized individual has a history of invalidation of his/her construing, particularly in regard to core aspects of selfconstruing. 2. This sometimes involves one or more episodes that lead to massive invalidation, and act as "transformative triggers." 3. The individual with a very undifferentiated (and thus inflexible) construct system may be particularly vulnerable to such invalidation and consequent structural collapse.		4. His/her radical beliefs, usually drawing upon available social constructions, allow the development of a "turning point" in his or her sense of identity with a more structured and certain view of the world. 5. The development of an extreme negative construction of another group, which may be perceived as responsible for the individual's invalidations, allows further definition of the self by contrast with this group. 6. The individual's radical constructions are validated by contact with others who share similar views, often coupled with constriction of their previous social world to avoid further invalidation.	7. The likelihood of acting upon radical beliefs, including violent actions, is greater in those individuals in whom beliefs in such actions provide the greatest increment in the structure of his/her view of the self. 8. Reconstruing of violence as acceptable may be necessary if the person is to engage in such acts without guilt (and indeed to experience guilt for not engaging in them). 9. His/her radical view of the world may be shored up by "hostility" in Kelly (1955) sense of extorting evidence for the individual's constructions. 10. Similar processes may operate in members of the "other" group, creating a vicious cycle of extreme construing based on mutual validation of extreme negative views of the other.		
Mechanisms of Political Radicalization Pathways Toward Terrorism [12]	1. Personal victimization 2. Political grievance	3. Joining a radical group – the slippery slope	4. Joining a radical group – the power of love	5. Extremity shift in like-minded groups 6. Extreme cohesion under isolation and threat	7. Competition for the same base of support 8. Competition with state power – condensation 9. Withingroup competition – fissioning	10. Jujitsu politics 11. Hate 12. Martyrdom
Counter-Messaging and graph [13]	1. Sensitivity			Phase 2: Group Membership		Stage 3: Action
A Path of Radicalization Complementary of the Group and the Individual Flaw [14]	1. Personal vulnerabilities			2. Psychological Functions of the Group. 3. Functioning Dynamics of the Radical Group.		
A Simulation Model of the Radicalisation Process Based on the IVEE Theoretical Framework [15]	"a person's individual vulnerability to radicalisation, their exposure to radicalising moral contexts, and the emergence of radicalising settings"					
The Three Ps of Radicalization: Push, Pull and Personal. A Systematic Scoping Review of the Scientific Evidence about Radicalization Into Violent Extremism	Push, pull, and personal factors. "push factors largely focus on structural, political, and sociological explanations; pull factors tend to focus on group-level sociocognitive explanations; and personal factors are concerned primarily with individual psychological and biographical explanations."					

Exposure to misinformation, conspiracy theories, and other radicalising content is increasingly easy for people to experience given the ubiquity of digital communication methods. Initial encounters with radicalising topics may come in a number of forms, including stumbling upon them in a search for resolutions to personal uncertainties, through social media algorithms, or via word of mouth from friends and family, either face-to-face or online. However, most people who encounter fringe content, which may even be on "a daily basis", do not go on to adopt extremist views themselves but rather reject extremist claims and narratives [10]. As El-Muhammady (2019) [10] notes, this capacity to resist misinformation and fringe views "is particularly true for those who have a strong value system, self-identity and intellectual capabilities"; traits that starkly oppose those above-mentioned personal characteristics that can leave one vulnerable. Thus, neither personal characteristics nor exposure to misinformation alone can be understood as a trigger for the radicalisation process—rather, both must co-occur for one to be drawn into this first stage.

In our data, trigger incidents were more frequently identified than potentially influential personal characteristics, however, the nature of these triggers typically provided insight into the person themselves. Among those who joined far-right communities, circumstances that contributed to radicalisation were a strong sense of loneliness and inadequacy due to either a lack of friends or negative experiences with girls, along with a chance encounter with some kind of role model and community into which they were welcomed; either online or face-to-face. These individuals were all male, based on the information disclosed, and typically identified their age of vulnerability as around 12–14 years of age. Among anti-vaxxers, such experiences ranged from distrust in their country's healthcare system exacerbated by the COVID-19 Pandemic, to the birth of a new child and a parent's anxiety about the child experiencing pain or hurt. The pool of anti-vaxxers included women and men for both of the above examples; as well as additional reasons including unknowingly joining a fringe community when seeking out parenting advice, to job insecurity. Among the three distinct groups examined in the development of this model, it was boys and young men who joined the far-right that had the most similar stories which depended heavily on personal characteristics; while radicalisation for anti-vaxxers typically required a stronger catalyst to trigger their adoption of fringe views.

For those who fell under the spell of the QAnon conspiracy, there were a range of topics that drew them to fringe communities; a trait that is perhaps reflective of the QAnon conspiracy's nature as an umbrella for a range of smaller conspiracies. These included the individual having a pre-existing interest in American politics and controversies like the Snowden leaks in 2013, or the experience of mental illness, which several posters felt had made them more vulnerable to being taken in. Several former QAnon believers identified themselves as former die-hard supporters of Bernie Sanders due to their strong distrust of governments—an element also present in the QAnon tale. These individuals had often been sent social media content such as Youtube videos about the QAnon conspiracy theory by friends who shared their views, which triggered their radicalisation. Many more identified themselves as having already held right-wing or conservative views, often due to their family's influence, and found these views were stoked by Fox News, the rhetoric of Donald Trump, and social media which gave them increasingly fringe content based on algorithms. Also unique to QAnon were identifications of religion, specifically Christianity, as a channel that made one susceptible to radicalisation due to the biblical nature of the QAnon narrative. Common to all posters was a kind of discontent in the state of the world or powerlessness in their own lives, as well as an inciting incident which caused them to seek out or come into contact with a source of misinformation and fringe views.

At this early stage of uncertainty, the common means by which individuals may exit the radicalisation pathway is access to reassuring advice from a trusted source. This could include information from a reliable source

such as a news report, a reputable website such as a government factsheet about vaccines, or consultation with authoritative figures such as doctors, community leaders, or even parents. It is crucial to emphasise that many people, especially people who spend a lot of time online, are routinely exposed to misinformation and experience uncertainties about topics that can fuel extreme views [10]. Therefore, many people can move into this first stage of radicalisation, particularly in the midst of periods of societal disorder, such as the COVID-19 Pandemic. However, most people easily resolve these doubts or dismiss the misinformation and do not progress further down this radicalisation pathway [10]. Increased access to, and more plain-language and appealingly-styled versions of these resources, which can be circulated on social media and hosted on government websites, would be a way to reach more people in this early stage.

2.3.2 Stage 2

Although not clearly marked as a distinct stage in existing literature, analysis of posts and interviews by now-recovered believers of misinformation and conspiracy revealed another common experience of the radicalisation process. This was a stage that saw increased curiosity in a topic and in some cases the beginnings of information-seeking behaviour to clarify one's thoughts on the topic of concern. For example, questioning the effectiveness of vaccinations to one's doctor or sharing tentative disclosures of interest in fringe views to existing social networks. In these accounts, experts and peers alike who reacted with a stigmatising reaction to the topic, like disbelief at the individual's views, or some kind of mockery or shaming, accelerated the radicalisation process for the vulnerable individual, who made a rapid retreat to a "safe" online space related to the topic for assurance.

It is speculated that one reason why this disclosure/stigmatisation stage may not be present in existing models is because many deal specifically with radicalisation related to violent religious extremism; a fringe perspective that is widely condemned due to high-profile terrorist attacks in recent decades. As a result, individuals would be highly aware that their newfound perspectives would likely not be welcomed by many people around them, and may more cagily guard their activities. In contrast, doubts or confusion about medicine, politics and other government activities, or daily uncertainties (e.g. the COVID-19 Pandemic) are topics that are often within the scope of ordinary conversation for many people. As tentatively raising one's views on these topics is not socially egregious or inherently criminal the way violent terrorism is, those who are drawn into conspiracy theories and misinformation may be more likely to experience this stage of the radicalisation process.

In the original data examined for this project, doctors, friends, and romantic partners were among the most-mentioned groups to whom vulnerable individuals disclosed their emerging fringe views. This disclosure was often a significant investment of trust for the individual, who had to genuinely work up the courage to share their thoughts and questions with others. As a result, reactions of judgement, disparagement, and aggression from peers and mentor-figures produced visceral shame and embarrassment for the individual, at times making them regret having placed their trust in that person. They then retreated to online spaces where they either read content that further affirmed their views, or in some cases, tentatively posted about their experience, for which they received support and encouragement from others further down the rabbit hole. This experience, while not universal, caused them to become further radicalised.

The emotional strength of the accounts in our sample about the impact of stigmatising reactions from peers and mentors provide a clear indication of what a deradicalising off-ramp would be at this stage: an

empathetic, non-judgemental reaction from the same cohort. In one example in our sample, an individual was able to leave their fringe community at this early stage due to an empathetic doctor. It is important to recognise that such peers and mentors in the sample were not identified as having been mean-spirited, but rather, having not fully recognised the sensitive nature of the issue. One method to perhaps encourage more understanding responses would be the creation of resources for people on how to talk to a friend or associate (e.g. a patient) who is expressing fringe views. These resources could guide peers and mentors to identify the underlying issue, and provide the affirmation that would otherwise be sourced from fringe online spaces, whilst also guiding the individual to a solution for their anxiety such as a trusted resource recommended by the friend.

2.3.3 Stage 3

At this stage of the radicalisation process, the individual has started to view themselves as part of a community centred around the issue. This internalisation of one's views on the topic starts to transform their own sense of identity, so that their 'in-group' allegiance is now more closely associated with their new fringe networks. This is recognised in existing models, such as by Winter and Feixas [11], who identify "the development of a 'turning point' in his or her sense of identity with a more structured and certain view of the world" as a key part of the process (p.4). Likewise, El-Muhammady [10] recognises this middle stage as one characterised by a "deepening" of the person's new views and investment in a new path of some kind (p.160). In our data, a common behaviour that enabled the deepening investment in the new community was more time spent online in related fringe spaces. Specifically, this often involved more reading and "research" into the topic, such as the threads of a specific QAnon theory or details about vaccine reactions. At this stage of deepening investment, many people reported still feeling grimly about the topic with which they were concerned, but gained an increasingly addictive thrill from "connecting the dots" between their own anxieties and plausible explanations online.

Another feature identified during this stage in our research is the withdrawal from existing networks, and the beginnings of perceiving those with different views as 'outsiders'. This can manifest in behaviours like the self-censorship of discussion about their topic(s) of concern with those existing peers, as in one example, as well as a withdrawal from mainstream information 28 sources. As Winter and Feixas [11] note in steps 5 and 6 of their model, and McCauley and Moskalenko note in stage 4 [12] in their model, increased negative perceptions of an opposing group as well as "constriction of their previous social world" also occur around this time. For some individuals we observed, they reflected that as their investment in their new views intensified, they began to behave less kindly towards people they knew who were "outsiders" to their views, such as women. Rather than casting this trio of changes as three different phases, in our model, we see these behaviours as co-occurring as the individual re-negotiates their own identity. As a result of their increased reliance on fringe community for social and informational needs, which validates their feelings, we also include in our model the emergence of feelings of happiness, certainty, hope, and reassurance. The alleviation of the earlier anxieties which drove them to this community is, in part, driven by the discovery of a new community they trust, who can explain away lingering doubts and who provide steps an individual can take to achieve the desired stability in their own life.

At this stage in our sample, when the individual was withdrawing from mainstream informational sources and their existing peer networks, there were fewer opportunities for traditional fact-checking methods to reach them. It is at this point when the value of organic de-radicalisation spaces becomes apparent. These

spaces, which provided the data for our research, allow radicalised individuals to speak to others who have previously shared their beliefs, giving them an opportunity to be understood whilst also being provided with more reliable information. In our sample, there were some instances where people stumbled on these spaces while caught in a conspiracy-related research spiral, and this triggered the realisation that they wanted to get out', but did not know how to begin. They were able to make posts to the space explaining their situation, and de-radicalised former believers gave advice on how the individual can leave the fringe community. Many users expressed immense shame at having been involved in a fringe community and did not want to discuss it with a therapist or friend for fear of admitting what they had believed; however, former believers were able to withhold judgement and gain the trust of new arrivals. There has been limited research on these spaces [18], but based on the posts there, it is clear that they offer a unique form of support. Finding ways to promote and support such spaces, either through algorithms or other arrangements, could increase their reach and effectiveness.

Another exit path that emerged at this stage was the individual limiting or leaving of social media and the related online community. This was done willingly by individuals, usually following the realisation that their investment in this community was worsening their insecurities rather than improving their life. For example, they were noticing that outcomes promised by their online community were not fulfilled, or that their mental and/or physical health was declining. The decrease in online involvement, which could be spent on hobbies or with friends/partners, gave them perspective on their lives and positive experiences of the world. This allowed them to re-evaluate the impact of the online community and its views on their life, their goals, and the world around them. Given that research has illustrated that time spent online and on social media is detrimental to human health, the impact of living a more analogue life should not come as a surprise. In terms of ways to trigger this exit path at this stage, more public awareness could be raised about the impacts of social media, and crucially, how to recognise excessive digital usage and "log off". This could be promoted online, as well as through school education to reach younger age groups to develop healthy digital habits at a younger age and reduce their exposure to misinformation in the first place.

2.3.4 Stage 4

At this stage of deep immersion in the fringe community and its beliefs, most research identifies features such as lifestyle change, change in peers, and change in worldview as key components. Winter and Feixas (2019) heavily emphasise the increased hostility of individuals from stages 7-10 of their own model, including a more derogatory view of out-group members and an increased acceptance of violence. van Eerten et al. [17], drawing on their use of the penultimate stage of Doosje et al.'s model([13]), identify the adoption of group membership as crucial to the increasing depth of the person's beliefs. The group identity, at this stage, provides a kind of "shield" against de-radicalisation efforts because the community itself is valued by the radicalised individual (p.29-30). Catherine and Louis [14] also draw attention to the role of a group's worldview and also their dynamics in these late stages of radicalisation, highlighting the significance of the newly acquired identity in keeping the individual radicalised. Several individuals in our data reflected that efforts to talk them out of their views on the basis of logic were highly ineffective—instead, they became more adept at defending their new views when they arose in conversation with their existing social networks. Our model also notes that the individual's reliance on the community for social needs and a sense of self has become a dependence at this stage, with individuals in our case studies reporting an increasingly black-and-white or us-vs-them mindset, as well as an increased comfort with the idea of violence in the name of their group in some instances. This was particularly noted in relation to women as well as perceived political opponents, with

both cast as conspirators who had to be brought under control at all costs. As a result, this animosity could emerge during encounters with their day-to-day peer network, no-longer concealed from view, and this could prompt increased stigma from existing networks and a further loss of support.

We also identified a number of features more broadly relevant to fringe communities that largely operate online. For example, spending large amounts of time online to one's own detriment, such as causing serious sleep deprivation, illness, and a loss of social and leisure time, is one factor that was identified in our case studies. Klausen et al. [9] also note this, however given that their model is largely concerned with extremism that can lead to terrorism, the role of excessive time spent online is placed earlier in their 'detachment' phase (p.72). Klausen et al. [9] also identifies increased hostility to others and making seemingly abrupt changes to one's lifestyle as traits of this stage, which in our model were mostly identified in instances when people felt they were most deeply immersed in their online radicalised space. The investment in this space could include some kind of active participation by this stage, such as contributing one's own interpretations of (mis) information, and sharing their own sources with others—again, a behavioural emphasis on information that is not as emphasised in other models we have discussed. For many, it was the consumption of misinformation and positioning of the self within the greater narrative they were consuming that provided a sense of agency and action, even if they were not actively preparing for violence themselves. However, one case study who had joined a far-right group and had an established offline community by this stage began bracing themselves for participation in violence by starting to acquire weapons, indicating that perhaps face-to-face connections can prompt one to some kind of action at an earlier stage.

As is described in Catherine and Louis [14], the new-found sense of identity, community, and source of guidance in one's life sometimes led to feelings of euphoria, and a strong sense of being among the "special" or "chosen" humans who had been called to a particular cause (p.190). This euphoric feeling could also be interspersed with periods of severe doubt, which were soothed by the new social network who offer reassurances and answers to these doubts. For at least one individual, the euphoria brought about by the deep involvement in QAnon supported them through periods of severe depression related to their life circumstances at the time. In our data, the role of the community as information sources had the dual effect of maintaining an individual's social ties to the group, while further restricting any remaining outside input of information which could otherwise challenge their worldview or open cracks through any doubts. In this insulated state, efforts to reach the radicalised individual through fact-based discussion also had to contend with the individual's new sense of identity and community ties, which provided a strong defence against outside influence. Likewise, at least one user actively sought to prevent members of their online community from disengaging from the community and abandoning its views, thus preserving the community and worldview.

We also include the individual's active withdrawal from existing social networks as another key manifestation of this stage. While the person may have already distanced themselves from weaker social links, at this later stage, we found that more proactive decisions were made to sever ties with friends and family. This could be due to the belief that the person was holding them back, untrustworthy, or even somehow dangerous to the radicalised individual. These decisions caused much grief for those who were cut off, according to former radicalised individuals who were now aiming to repair those connections.
At this stage, some of the previously-discussed exits from the radicalisation pathway still benefited individuals, such as a trusted, empathetic friend, limiting social media use, or the organic deradicalisation spaces. However, at this higher level of investment in the fringe community and active withdrawal from existing

networks, we also saw the emergence of several new exit pathways. The first of two new forms were the experience of a personal shock that forced them to reconsider their views. This could be the death of a family member or friend, a significant personal realisation (e.g. sexual orientation), or something related to the fringe community itself, such as election outcomes. These experiences were so deeply personal that the individual was forced to confront their views in order to grapple with the new circumstances of their life. As a result of this need to reconcile their new reality with the alternative reality provided by their fringe community, many of the individuals in our sample were able to begin withdrawing from those spaces and make efforts to rebuild their lives. This indicates that intervention methods at this stage need to be able to reach an individual at a deeply personal level, rather than relying on facts, debate, and logic. The second new exit pathway that emerged at this stage was, in contrast, something that was highly detached from the individual. Being exposed to third-party content that exposed the falsehoods of a fringe view was a means by which several people were prompted to re-evaluate their views. This could come in the form of authoritative documentaries about other conspiracies or scams, as well as user-generated content that impartially highlighted the illogic of a fringe view. Crucially, none of these were government-made; rather, it was the sense that the content was by a party not especially invested in a particular view and was simply noting the truth that reached these people. This raises possibilities around whether algorithms on digital media platforms could be harnessed to help promote this content, as well as relevant documentaries being made more available. The value of this content was in its revelation of how conspiracies, scams, and misinformation can be created and promoted, from an impartial stance. While direct fact-checking merely made people feel under attack and defensive, with this content, individuals could be reached as they had left their guards down.

2.3.5 Stage 5

At stage five, existing models and our own show clearer signs of divergence. In common with ours is the view that, at a second-to-last stage, individuals are feeling an increasing sense of need to 'give back' to the group they are a part of, or 'take action on behalf' of the group's goals. This stage of heightened motivation and even preparation is understood by Klausen et al. [9] to be a stage when one involved in religious radicalisation may undergo training with peers, to learn how to carry out violent extremist attacks (p.72). For El-Muhammady (2019), this stage is similarly defined by some kind of process by which the individual is "externalising" their views, which might include working together with others (p.174). At this stage, therefore, the individual's radicalisation state is now visible to others as more than a passing eccentricity.

Physical manifestations of erratic views in the individual's own lifestyle are more broadly defined in our model, to reflect the wider range of radicalisation experiences we observed. For example, rather than needing to flee overseas for wartime training, the internet allows those who wish to seek out such skills and materials to source them online. Furthermore, the existence of digital platforms allows for potential participation in public displays of social division that do not require training or violence. For example, participating in hashtag campaigns to trend their own issues in mainstream social media spaces, to potentially recruit new members. Likewise, social disruption that did not cause physical harm to another person but which somehow expressed their views was one example in our data, in the form of a targeted act of vandalism.

At this stage, the two exit paths in our sample were highly similar in nature. The first was medical intervention, often brought about by an existing issue, such as for anxiety, depression, ADHD, or addiction. Posters did not give context for whether they sought this help themselves or whether they were urged to do so by people around them, but they did reflect on how addressing the severe personal health issues they were dealing

with had triggered their deradicalisation process. As their wellbeing stabilised, they were able to take stock of their life, and battle their dependency on coping mechanisms, which includes the fringe communities and worldviews. While not every radicalised person necessarily has health problems, or the means to seek professional support for health problems, this does indicate an alternative route by which authoritative sources such as medical professionals can enter the lives of these individuals as a trusted source of support. The second means by which deradicalisation was triggered at this stage was one example of a person being forcibly institutionalised for mental health and related issues, after a severe breakdown. Again, while this is not necessarily applicable to all radicalised people, it reaffirms the level of effort and resources required to reach someone at this stage. From these examples we can observe the importance of providing a new structure or plan for the future for the individual, to help with the loss of the fringe community and views they depended upon. Creating clearer channels that are accessible to the public, such as online resources and support groups, to help families help loved ones leave these communities could provide more opportunities for stronger interventions.

2.3.6 Stage 6

Overt extremist acts were uncommon in the data we examined, likely due to the posters having parted ways with their fringe community and much of the associated views. However, numerous examples of people who have gone on to commit these acts are available in news media, ranging from QAnon supporters storming the US Capitol building, to more individualised actions such as the Wieambilla police shootings in Queensland in December 2022. Our model offers a broader recognition of extreme acts in line with the types of radicalisation observed in the project. It is of course acknowledged that there are potential acts that users could commit (or have committed) that they would choose not to disclose on a Reddit forum. Therefore, a stage six extreme act in this model could anything which may negatively materially impact someone, including solely themselves, or others; and which is completed with intention and ideological belief.

In contrast, in the models considered in existing literature, there is a common understanding that the final stage of the radicalisation process entails some kind of violent extremist act, such as a suicide bombing. The focus on violent acts in existing models is likely due to the original scope of these projects having a focus on violent and religious extremism. While the models themselves are appropriate to their scenarios, as a result of this, they cannot be used to understand the newer kinds of radicalised acts that are taking place in an age where misinformation and conspiracy are so easily transmitted across digital spaces and apply to so many topics. For example, in our dataset, the most extreme self-confessed act was among anti-vaxxers, who decided to stop vaccinating their children. Given the considerable threat this poses to the life of a young child even under normal circumstances, to say nothing of the risk posed by COVID-19, this decision constitutes an extreme act within the context of an anti-vaxxer community. In the context of other communities in our dataset, which were the far-right and QAnon, there was a greater appetite for violence revealed in earlier stages, but no self-disclosed acts in these views. This pattern support's Kruglanski et al.'s observation that most radicalised individuals do not go on to commit extreme acts (2022). Rather, in the framing of our model, these individuals would hover somewhere in the middle stages of the model—stages four or five—where they fuel the online culture of their fringe community, but do not act on their fringe views in a manner that causes direct material harm. At stage six, the primary exit path would likely be a deradicalisation program such as one organised by a government organisation, as has been demonstrated in previous academic studies [10].

2.4 Application of pathways to personae

This section details a likely radicalisation pathway for each of the three personae defined in Section 1.1.1. Similarly to the construction of personae, these journeys are fully fictitious (not based on a real case), but exemplify an expected evolution of a person associated with a given persona.

2.4.1 The "Jennifer" Persona

Stage 1. Jennifer has existing dissatisfactions with herself as a mother and whether she is still an attractive woman after having kids; experiences high pressure due to taking care of children, and misses her friends because she can't easily socialize with children. While browsing social media as her replacement for social activity, she sees comments in a mother's group about not wanting to have her children vaccinated for preschool. This catches Jennifer's eye, and she starts to wonder if she should also be worried about this.

Stage 2. Jennifer's curiosity and uncertainty in the potential concerns around vaccinating her kids leads her to read more of the articles those mothers shared, and even messages some of them to ask for more details, which they kindly provide. She mentions it to her husband, who brushes it off as "fine" without a full conversation. During a rare catch-up with her friends, she also tentatively mentions her concerns, but they roll their eyes and say she can't honestly believe that. Feeling humiliated, Jennifer messages some of the mothers soon after, who are appalled at her friends' behaviour, and explain that they just don't understand because they don't have kids. But Jennifer needed worry about them, they have their own private facebook group she can join, and it's a safe space. (Mother's Voice, Mother's Choice)

Stage 3. Jennifer resolves to not talk about her anxieties about vaccination with her friends or husband again. She starts ignoring news articles and mother's advice brochures on the topic, which she feels don't answer the questions she has. She spends more time online bonding with the new mother's group and reading the articles they link too, and loses interest in her previous activities of yoga and painting. She grows more worried about vaccinating her kids.

Stage 4. Jennifer starts declining her friend's invitations to catch up, to their confusion and hurt, and often spends dinner with her husband scrolling through long comment threads about vaccine dangers on her phone until he tells her with frustration to put it away. While he's at work, she's started writing her own posts for the group, based on her googling into the various ways she might legally prevent her children from getting their vaccinations. She's considering homeschooling, but hasn't raised it with her husband, who is hassling her to resume her painting.

Stage 5. When Jennifer's friends message her with their concern that she has seemed unhappy with them and offering to visit her at home to catch up, Jennifer is embarrassed at the thought of them seeing her having put on some weight, and angry that they are blaming her for not being able to catch up. She tells them not to bother her again because they don't understand what she is dealing with as a mother. She has raised her preference to not send the kids to preschool with her husband and instead homeschool them, but he disagrees that this is the best approach, angering her. When he suggests that she seems stressed and should consider resuming her yoga, she feels he's called her ugly, and the pair have a huge argument in which she tells him he doesn't understand how hard it is to be a parent because he spends all his time at work. When he

moves her laptop later to try and tidy the house a bit more, he sees that she's been posting her complaints about him and her friends to an anonymous twitter account, where she's also joined in on anti-vaccination hashtags.

Stage 6. Jennifer learns her husband saw her online activities, leading to an even more extreme argument. Jennifer messages her mother's group in despair for advice. They tell her neither she nor the kids are safe with him, and the next day while he's at work, she writes him a letter asking for a divorce. She leaves with the kids and takes them to the house of a nearby women from her mother's group who offered to let her stay with some other, similarly-lost mothers. In exchange, Jennifer will work for free as the communication specialist for the group's social media and website, as well as a petition they are launching to send to a politician about the urgency of making vaccination of children optional.

2.4.2 The "Patrick" Persona

Stage 1. Patrick has felt increasingly lonely in the years since he retired from being a primary school teacher. Before COVID-19, he volunteered at the local library to help children learn to read, but that program has since been cancelled. He now spends most of his time at home with his wife, who has become very frail with dementia and requires full-time care. He also experiences ongoing back pain, from helping her move around. The highlight of his day is when their grandchildren visit for a few hours after school and fill the house with energy, and he helps them with their homework. His daughter signed him up to Facebook because she thought he could talk to friends there, and he uses it every night before bed. He has recently become intrigued by news articles that are shared online, about topics he never hears about on television; especially relating to child safety at school. Specifically, he's concerned that more schools seem to have sexual education at younger ages, which he doesn't think is needed.

Stage 2. Patrick's concern about the appropriateness of school education leads him to investigate the topic through several news sites and Facebook, where he encounters the page of a known Australian politician who is adamant that children are being taught sexually explicitly material in schools for malicious purposes. Patrick knows not to believe everything on the internet, but he also knows the poster is an actual politician, which gives him pause. Out of concern, he tries to ask his daughter if she knows anything about this, but she just looks worried and offers to take him to the doctor. His wife doesn't really understand when he tries to tell her, but says she trusts he's right. He decides to follow the politician online, and sees in the comments that people often complain about the state of school education or people dismissing their concerns. This reassures him that he's still in control of his thoughts, and what's more, that he should be allowed to express them.

Stage 3. When Patrick's daughter later asks how he is feeling, clearly remembering his strange question about school health education, he reassures her that everything is fine. Internally, he is wary of his daughter's views, as he's noticed she sometimes 'likes' posts about youth social movements, which he thinks are a distraction from their real learning. He's begun to notice that the news pays more attention to those topics than the issues of child safety he's increasingly concerned about. He wants to do his own research. Links posted in the comments of the politician's page, which he now clicks on to read and consider, suggest that there is more to worry about than a sexually explicit school program. In fact, this may be the tip of the iceberg. He spends more time at night reading about alleged networks of child traffickers, until his wife falls asleep beside him and he guiltily gets her ready for bed, late yet again.

Stage 4. One afternoon, his wife mentions to their daughter that he doesn't read to her in the mornings anymore because he spends all his time on the computer. She confronts her father and tells him to take better care of her mother, and he says nothing but thinks she should take better care of her kids. He is spending more time on new forums he found through other commentators online, where there is even more evidence about these child trafficking groups. He's also realized he has a knack for interpreting the evidence and making connections, and everyone is always amazed at how well he can analyse things. After several years with only his ill wife and grandkids to speak to, it's a relief to finally use his brain more, as well as speak to people more often.

Stage 5. When one of his grandkids enters year 7, and receives homework that asks her to prepare a short report on "the ovulation cycle", Patrick becomes alarmed and demands to know which teacher gave her this work. She becomes frightened and calls her mother, who arrives and takes the kids away immediately. As she goes, he angrily tells her about how the report is part of a testing process to determine which girls should be harvested for eggs to put into beauty products for celebrities. She tells him she'll come to take him to the doctor tomorrow, and he tells her not to come back at all if she doesn't believe him. When she goes, he rings the school of his grandkids and demands to speak to the principal, but they hang up on him. His wife manages to calm him down by calling him over because even she can tell something is wrong. After he helps her to bed that night, he goes online and tells everything to the people in the forums, including his fear for his grandkids and regret of not explaining it properly to his daughter, as well as his fury at their school, which he now knows is part of these trafficking rings they have been investigating.

Stage 6. Patrick knows that neither his grandkids nor any of the other children at that school, will be safe unless he does something to help them. His daughter has given up on trying to take him to the doctor and instead moved her mother out to live with her. She had been getting thin, because he has been so busy planning he sometimes forgot meals, and she could never remind him. One bright Monday morning, he dresses in his best suit and attends an appointment at the school on the pretense of applying to host an after-school reading club. During the interview he attacks and kills the Principal, severely injuring several administration staff who attempt to subdue him and the police arrive.

2.4.3 The "Aaron" Persona

Stage 1. Aaron is an ambitious student in his final year of law school, and everyone assures him that his destined for a great career. Despite this and his luck at landing an excellent role as a summer clerk at a prestigious firm, he feels inadequate compared to some of his mates at his Church group. Some have long-term girlfriends and intend to be married soon, and some have been employed since they left school at 16 to pursue a trade, and they tease him about spending so much time studying. He is aware of his health and tries to find time to go to the gym, but he is short on time, and often relies on skipping meals to stay at his health app's calorie goal, often with a big burger meal or pizza to get the protein goal in particular. He is frustrated that he still doesn't really look fit, and starts spending his spare time on public transport searching for workout tips. He decides to invest in a tub of protein powder, and starts researching the best brands. He finds some interesting articles about the need for high levels of testosterone to improve strength and bookmarks these to read later.

Stage 2. When his summer job ends and his final year of studies truly begins, Aaron finds he has a bit more time to go to the gym, but he doesn't feel any fitter. He decides to look more into the stuff he saw on the importance of testosterone, which leads him to a popular men's fitness and lifestyle blog that frequently

discusses it, along with the value of steroids for being fit. His mother asks him about the protein powder, and when he explains its purpose, she tells him he shouldn't change anything about himself. This annoys him, as that's exactly what he wants to do. He snaps that he's also been looking into steroids because they "aren't that bad"—mostly because he knows it will annoy her, and that makes her roll her eyes and ask him to get his father for dinner. Feeling irritated after dinner, he sees that the blogger he follows is doing a live video and decides to watch it to feel better. The blogger talks about how others might try to hold them back from their goal to be strong on their own terms, and everyone in the comments agrees. Aaron writes a quick comment agreeing, and saying his mother is his big problem, and feels better when loads of other guys say the same thing. They also link him to some other forums saying he should come along and talk there, since plenty of them feel this way.

Stage 3. Aaron starts to really spend time on the forums he discovered. He's very interested in several threads that specifically discuss ways to naturally raise one's testosterone, which should be more effective than protein powder, and also more natural. A few suggestions also suggest trying to avoid processed foods, and although he loves junk food, he swears he'll give it up. He is also glad to know that many of them are Christian like him, but unlike the guys at Church who seem to have everything figured out, these guys also have similar problems with fitness and girls. However, they think he's awesome for studying to be a lawyer and it's the first time he actually feels truly proud of how hard he works. Because of his late nights, split between studying and being online, his sleep cycle is disrupted and he misses a few Friday evenings and Sunday mornings at Church. Although his mother notices this, she puts it down to university stress. After a few curt replies to her reminders, he does try to make an effort to go back, but when he sees his mates, he doesn't really feel like he's part of the group anymore.

Stage 4. One night at the Church group, one of the guys makes yet another joke about Aaron still being single and looking too skinny, and he feels like he can't stand it anymore. He storms out, and when he gets home, he writes about it on the forum. Everyone agrees the group is full of losers anyway, and some of the guys are also eager to hear that he has actually been losing weight since cutting out processed food. They collectively agree that he might get even better results if he switched to all raw foods, which he says he'll do and report back on. Several other guys also agree. Aaron's parents find his sudden switch to raw foods only concerning, especially his insistence on raw eggs, but he ignores them. He also starts to wonder about whether water can be processed. When he asks the guys online, many say they only drink bottled water because the fluoride in tap water suppresses testosterone. He looks into the history of fluoride in Australia and learns some areas voted against it and don't have it. He posts this information to the forums, with the point that if some people are able to opt out, then why not let everyone have that choice?

Stage 5. To his parents' distress, Aaron refuses to go to Church anymore, and spends much more time online. He has become very thin and often feels unwell because of his extreme diet, and is spending much of his savings on bottled water. Many of his university friends have also started avoiding him in classes, and unfriending him on social media because his posts about the danger of tap water have started to annoy them. His friends online remind him that he doesn't need them; what he needs to do is find a way to draw attention to how dangerous their water is. He is also increasingly angry that he has done all this work to try and attract a girlfriend, but now the girls in his classes won't even look at him. He begins to feel angrier at them too, for not recognizing all the effort he's put into becoming attractive enough for them. When a girl in class laughs at a mistake he makes in answering his lecturer's question, he flies into a rage and verbally abuses her until security escorts him away. He gets a notice of an act of university misconduct and vents online.

Stage 6. Aaron and his friends make a plan, which begins with him accepting punishment for his 'misconduct', and volunteering to help with the next university social events as community service. For the public fair organized for International Women's Day, he offers to help with the food and drinks set up. During this time, several friends from online arrive including one who works has stolen toxic chemicals from his workplace. The group add it to the catering food being set up, as well as the premade drinks, leading to several hundred people being poisoned or becoming severely ill.

2.5 Application of pathways to public case studies

In this section we apply the above-discussed pathway to public examples collected from news articles in which former misinformation-believers have been interviewed.

2.5.1 Lydia's story

[19][20][21][22] **Stage 1.** In 2008, Lydia took her first child to have her first vaccinations. She was "nervous" about this, and her fears seemed to be confirmed when her daughter "screamed and cried" in response. When her daughter was still crying later, and experiencing a minor reaction, she rang a nurse and expressed her view that something was wrong. The nurse told Lydia that her daughter's reaction was normal, and that her concern was natural for a first-time mother. This did not reassure Lydia, but rather left her feeling "very brushed off … written off." She decided to find out more on her own, and began searching for information online. Lydia found an online forum for mothers, where she posted about her daughter's reaction. The other users responded with a range of possible explanations for her daughter's distress at the vaccine, including suggesting that it had caused her brain to swell. She described the online community as people who "…give you an answer that the other people couldn't give you or didn't give you. And so now you don't have any trust because you get an answer, right? You kind of run with that." Suddenly, Lydia had an answer for her daughter's reaction, and decided she would only allow her daughter to get certain vaccinations in the future.

Stage 2. Lydia developed more fears about vaccines, and in continuing her reading online, she was able to find a terrifying answer for every question. When she expressed her anti-vaccine stance to doctors, she recalls them "reacting with vitriol" at her views. She added that this reaction "just made me close myself off further – I felt really judged and upset and hurt and embarrassed." Her husband, who was indifferent about vaccines, did not try to influence her perspective either way.

Stage 3. Lydia's further immersion in the online anti-vaccination community ultimately led her to discontinue all vaccines for her daughter, and her next two children. The decision to not pursue vaccines was further made appealing by the fact that it did not require her to do anything, thereby excusing her from any responsibility for the outcome. She developed the view that organic treatments would help to boost natural immunity, and were superior to medicine.

Stage 4. When Lydia's first child started preschool, she surrounded herself with anti-vaccination mothers, giving her an offline community that held the same views as her online community. She avoided discussing vaccines with her family or anyone who supported them, to shield herself from being judged. She also recalls that, among her own community of middle-class mothers, there was the sense that having the ability to invest

in other health options meant she didn't need to bother with vaccines–they were, in her own words, "for poor people".

Lydia's exit. In 2020, when the COVID-19 Pandemic commenced, the shock of it forced Lydia to rethink her views on vaccines. She observed that a lot of money was being poured into developing them, and tentatively began reading up about vaccines from more authoritative sources. She learned about how preventable diseases like the measles were coming back due to vaccine hesitation, and realised that her own, unvaccinated son was autistic, and therefore his condition couldn't have been a vaccine reaction. This process of realisation and re-evaluation was painful for her as she had invested 12 years in her anti-vaccination beliefs. But, due to her fears about COVID-19, she anxiously had her daughter, then her two other children, catch up on their regular childhood vaccinations. She has since formed the organisation Back to the Vax (https://backtothevax.com/) with another former anti-vaccination mother to try and reach mothers who have been misled, and is studying to become a public health nurse to reassure new parents who have the fears she did.

2.5.2 Megan's story

[23] **Stage 1.** Megan was an extroverted young woman who had hoped that Bernie Sanders would become the United States president in the 2016 election. She had held a strong dislike of Donald Trump, considering him to be "racist, sexist and a Hitler-wannabe." However, in 2020, amidst the COVID-19 Pandemic and the expansion of the lockdowns in June, Megan was unable to see her friends or get space from her depressed and overworked fiancé. One night, after a day of fights with him, a friend sent her a link on Whatsapp to a Youtube series called Fall of Cabal. It told the story of the QAnon conspiracy, of the corrupt deep state, and how Donald Trump would save the United States. She watched all 10 episodes that night. It seemed to confirm all the suspicions she had of billionaires and the US government, which had originally made her inclined to support Bernie Sanders. Uniquely, compared to Sanders' failure to be elected President, the QAnon conspiracy offered her hope that someone in power was working to improve the world.

Stage 2. Megan had been converted overnight, and her fiancé noticed an instant change in her. She was "beaming and cheerful", but she initially was reluctant to explain why to her fiancé. When she finally told him her new views, he was disgusted and feared she might hurt him. Megan believed she could no longer trust him and spent more time online. She even prepared to move out and end their relationship, but her fiancé consulted with his therapist on how to proceed. With this support, and a boundary about when Megan was allowed to speak about QAnon with him, the relationship was stabilised. He discouraged her from posting about QAnon online, but a few times she did, and some friends insulted her and tried to convince her fiancé to leave her. She recalls that this pushed her deeper into QAnon, "finding solace in the community of like-minded people with whom I had a shared reality."

Stage 3. Megan's initial period of believing in QAnon was a kind of "mystical state or euphoria". She found herself praying more often, expressing her gratitude for the existence of Donald Trump, whom she now adored. Megan and her fiancé continued living together, but she had a limit on how much time she could spend talking about it with him. He reacted with calm further questions about what she said, and encouraged her to think about things more deeply without judgement. This patience caused some of her belief to develop cracks.

Stage 4. The euphoria of Megan's early belief wore off. She was left with only the underlying fears in QAnon: that the world was being controlled by evil, or that President Biden would mandate COVID-19 vaccines will malicious intentions. When she needed to, she was able to speak to her fiancé, who told her they needed to wait for more evidence about the vaccines, but agreed they could move to another country if a mandate was enforced. The pair reached compromises on issues like wearing masks.

Megan's exit. Megan's fiancé drew her attention to her now-depressed state, and how it was being caused directly by the time she spent on QAnon forums reading about world horrors. He convinced her to spend less time online, and focus on her wellbeing. Because of his prompts to think deeply about what she was reading, she had also begun to notice that various QAnon promises had not come true, such as the much-anticipated arrest of Hillary Clinton in relation to Pizzagate accusations. When the United States Capitol Building was attacked by rioters in 2021, the shock of hearing that KKK members were involved finally broke her belief in QAnon. She feels more positively about the future, and no longer identifies with any extreme politics.

2.5.3 Jadeja's Story

[24][25] **Stage 1. J**adeja was battling undiagnosed bipolar disorder and depression in 2016 before of Donald Trump's election, when he encountered Pizzagate conspiracy theories on Reddit. The conspiracy theories interested him, despite living in Sydney Australia, and having supported Bernie Sanders initially in 2016.

Stage 2. Jadeja was drawn further into these views when the first QAnon posts began the following year. The information he read was convincing because it gave an "explanation that, while it doesn't make sense, if it were true explains the situation better than the current explanations I'm getting" about issues of concern.

Stage 3. Jadeja enjoyed participating in Pizzagate and QAnon conspiracies, describing it as a form of "drug" where everyone contributed to piecing together the truth. He did not pay attention to fact-checking materials. He shared QAnon with his father, who also became deeply invested in the conspiracy.

Stage 4. Jadeja was spending "all day, every day for months" trying to consume more QAnon content to get the satisfaction of a new connection between the dots; another "hit". He adopted extreme views such as believing that the Former German Chancellor Angela Merkel was the biological daughter of Adolf Hitler.

Jadeja's exit. After two and a half years, Jadeja started to notice inconsistencies in QAnon claims. One day, when Donald Trump said a specific phrase alleged to be a secret keyword, and QAnon fans were celebrating, Jadeja encountered other information online that proved this phrase was commonly used by Trump. This shattered his beliefs completely: "It felt like in the space of one second, the entire universe collapsed in on me." He now publicly advocates against QAnon, and is trying to help free his father. Below are five other examples from online news articles in which former misinformation believers are interviewed, which are mapped against our pathway.[25][26][27][28][29]

Figure 2.2: Five other former misinformation believers' journey mapped against our pathway

HA Project Path	Category	1. Existing dissatisfactions with self, lifestyle pressures, lack of community, and general uncertainties lead to casual encounters with concerning topics	2. Increased curiosity in encountered topics, potential tentative disclosure to existing peers and stigmatised reaction (disbelief/mockery/shaming), retreat to "safe" online community for assurance	3. Self-censorship of topic discussion with existing social networks, withdrawal from mainstream information sources, increased reliance on fringe community for social and informational needs. Feelings of happiness, certainty, hope, and reassurance etc emerge, alleviating earlier anxieties.
Heather	Anti-vaxxer	Began to distrust vaccines when someone told her the flu shots had made someone else very ill.	The seed of distrust of vaccines grew until she became against them. She poste these views on Facebook and received angry responses, which confirmed she was doing the right thing. However, she also got positive responses.	Simpson was happy at how well her posts spread, and they were more effective the angrier she wrote. She was shocked out of her views when COVID-19 struck and she consulted with doctors.
Ashley	Qanon	When the COVID-19 Pandemic hit, she started to spent more time online, and Tik Tok began promoting conspiracy videos to her.	She started to spend more time online, and her young daughter noticed something was wrong.	Her time online expanded to Facebook, YouTube, and Telegram, where she spent hours every night reading conspiracies. Even after Joe Biden was elected president, she still believed in the plan.
Ivan	Qanon	Battling anxiety and depression when he encountered Pizzagate theories in 2016 and enjoyed reading them.	Ivan knew not to mention his views to others because they were extreme.	Ivan did not respond to fact-checking, considering it "boring". He considered Pizzagate/Qanon theories to be a kind of "team sport" activity and form of empowerment.
Craig	Anti-vaxxer	Craig was inested in wellness and health, even moreso after the birth of his daughter. Yet he became overwhelmed with the fear he couldn't protect his daughter from the world. He became convinced vaccines created autism.	Craig was provided with copious research and support from doctors, but could not be talked out of his views	Craig did not experience a period of euphoria, but fear.
Rein	Qanon	Rein was convinced the COVID-19 Pandemic was a conspiracy by comparisons to the Holocaust. She is the daughter of Holocaust survivors and this made her believe something sinister was ocurring.	Rein became interested in Qanon.	Rein's initial experience was of the optimistic and hopeful elements of Qanon about how the world could be saved.

Figure 2.2: Five other former misinformation believers' journey mapped against our pathway (continued)

HA Project Path	Category	4. Reliance on fringe community becomes dependence, involvement becomes investment through active participation, such as contributing own interpretations and sharing own sources, increased stigma from existing networks. Feelings of certainty become feelings of euphoria, sense of being special or chosen. In-group becomes a part of their identity.	5. Active withdrawal from existing social networks, physical manifestations of erratic views in own lifestyle, potential participation in public displays of social division. Increasing sense of need to 'give back' to group, or 'take action on behalf' of the group goals.	6. Overt extremist acts, which may impact only self or others
Heather	Anti-vaxxer			
Ashley	Qanon	When Joe Biden was innagurated, she rang her mother in a blind panic fearing the Democrats would steal her daughter. The failure of this QAnon "plan" led her to doubt the whole theory		
Ivan	Qanon			
Craig	Anti-vaxxer	Craig clung to his views like a religion, because the risk of being wrong was so severe. He finally reevaluated his views when he got divorced and is now a strong supporter of vaccines.		
Rein	Qanon	Rein's later experience about Qanon was about the certainty that a second Holocaust was coming.	Rein had a public meltdown attacking a standof facemasks that was filmed and sent viral online. Her family sent her to a psychiatric ward for treatment, which helped her to recover.	

3. Effectiveness of EU's Digital Services Act

In 2022, the European Union introduced the Digital Services Act (DSA), a new legislation to report and moderate illegal content from online social networks. Trusted flaggers are mandated to identify illegal content, which platforms must remove within a set delay (currently 24 hours). In this chapter, we analyze the likely effectiveness of EU-mandated mechanisms for regulating highly viral online content with short half-lives. We deploy self-exciting point processes to determine the relationship between the regulated moderation delay and the likely harm reduction achieved. We find that harm reduction is achievable for the most illegal content, even for fast-paced platforms such as Twitter. Our method estimates moderation effectiveness for a given platform and provides a rule of thumb for selecting content for investigation and flagging, managing flaggers' workload.

Misinformation, Disinformation and Illegal content. The Digital Services Act does not define what is illegal online. It only creates new EU-wide rules that cover the detection, flagging and removal of illegal content and a new risk assessment framework for very large online platforms and search engines on how illegal content spreads on their service. What constitutes illegal content is defined in other laws at the EU level or national level – for example, terrorist content or child sexual abuse material, or illegal hate speech is defined at the EU level. Where content is illegal only in a given Member State, as a general rule, it should only be removed in the territory where it is illegal.

However, the legislation mentions disinformation, election manipulation, cyber violence against women, or harm to minors online. In the EU, spreading false or misleading information is not generally illegal. Freedom of expression includes the right to express incorrect views. As a result, we conclude that the DSA legislation can be used to fight online disinformation (spreading false information with a particular purpose, such as foreign interference in political processes). However, it cannot readily address online misinformation.

3.1 Introduction

Social media platforms are the new town squares [30] – dematerialized, digital and unregulated town squares. In 2022, Elon Musk acquired Twitter with the stated goal of preserving free speech for the future. However, alongside free speech, illegal content disseminates and prospers in this unregulated space: disinformation that spreads faster than its debunking [31], social bots that infiltrate political processes [32], hate speech against women, immigrants and minorities [33] or viral challenges that put teens' lives at risk. In response, there have been calls for the governments to intervene and regulate. As the first move of its kind, the European Council introduced the Digital Services Act (DSA) and the Digital Markets Act (DMA) [34], EU legislation aimed at projecting the regulations of our offline world onto the digital one. It implements notice and action mechanisms (cf. Art. 16) to report illegal online content. Furthermore, the regulation introduces a process for appointing trusted flaggers, subject matter experts in detecting illegal content (cf. Art. 22). Once such content is flagged, platforms must promptly remove the content. However, online content is notorious for its "virality" – it spreads at high speeds and has short lifespans. Therefore, we ask about the effectiveness of this new legislation: **How to quantify the likely harm caused by illegal content, and how to determine the response time for effective mitigation?**

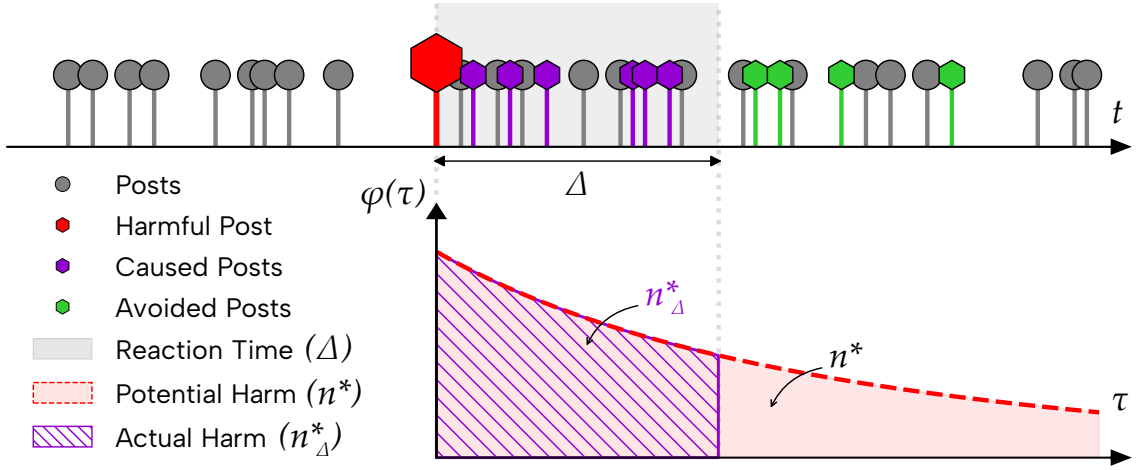


Figure 3.1: Social Media Dynamics as Self-Exciting Point Process. Social media posts (●) include a percentage of posts considered illegal (⬡). One illegal post (⬢) likely generates n^* other illegal content at the rate $\phi(\tau)$. When the post is removed at time Δ, the likely caused harm is limited to n^*_Δ (⬢) and further harm is avoided (⬢). The harm reduction χ is the percentage of all illegal content generated directly or indirectly via self-excitation avoided via moderation.

In this work, we leverage state-of-the-art information spread modelling to assess the effectiveness of the DSA regulation and the EU code of conduct for countering illegal online speech. Fig. 3.1 conceptualizes an online discussion, where each post (● or ⬡) draws more people into the discussion and generates more posts, referred to as offspring. This phenomenon of content spreading is known as the *self-exciting property*. A illegal post (⬢) will therefore generate potentially other illegal posts (⬢ and ⬢) with a decreasing intensity, shown by the red dashed line on the bottom panel of Fig. 3.1.

How would the new EU legislation potentially stop the propagation of the harm? The core concept is to limit illegal posts' reach and the offspring generation. We denote the number of illegal, direct offspring as the potential harm – denoted as n^* and comparable in meaning to R_0, the basic reproduction number of infectious diseases [35]. Content moderation is achieved by removing the illegal post (⬢) at time Δ after posting and thus stemming offspring generation after this time (⬢). In addition, we assume that any illegal direct offspring generated before Δ (⬢) are also moderated; their number defines the actual harm – labeled as n^*_Δ. The harm reduction χ is the percentage of all illegal offspring avoided, both direct and indirect – i.e., offspring of the offspring generated via the recurrent branching process.

Table 3.1: Variables of Interest for Modeling Content Removal

Parameter	Interpretation
$\phi(\tau)$	The rate at which content generates reactions on social media.
n^*	**Potential harm** – The number of additional illegal posts a content generates directly.
Δ	**Reaction time** – Mandated time to remove flagged illegal content on social media platforms.
n^*_Δ	**Actual harm** – The number of direct illegal reactions a content generates prior to moderation at time Δ.
$\tau_{1/2}$	**Content half-life** – Time until a content generated half the direct reactions.
χ	**Harm reduction** – Percentage of direct and indirect illegal offspring avoided by content moderation.

The effect of the policy heavily depends on the speed at which the discussions unfold on social networks. We quantify this using the content half-life, defined as the time required to generate half of the direct offspring. A recent (as of 2023) empirical investigation [36] determined the half-life of social media posts on different platforms: Twitter (24 mins), Facebook (105 mins), Instagram (20 hours), LinkedIn (24 hours), YouTube (8.8 days), and Pinterest (3.75 months). A lower half-life means that most harm happens right after the content is posted, and content moderation needs to be performed quickly to be effective.

3.2 Heavy Tails in Social Media User Activity

The initial work by Barabási [37], which analyzed email user traffic and discovered power-law distributed inter-event times of human activity, has inspired scholars to investigate the empirical inter-event time distributions of various phenomena. Extensive research on social media platforms has revealed power-law distribution characteristics in content-sharing dynamics. Lerman and Ghosh [38] examined the network structure of platforms like Digg and Twitter, highlighting power-law distributed user activity. Xiong and Liu [39] explored the opinion formation process and also observed power-law distributed user activity in their findings. Additionally, Bild et al. [40] investigated aggregate user behaviour and analyzed the systematic properties of retweet graphs, reporting a power-law distribution in Twitter inter-event intervals with an exponential cut-off. Similar findings were reported by Mathews et al. [41], who attributed the distribution to the burstiness of human behaviour. Notable early work by Crane and Sornette [42] focused on measuring the response function of social systems. They observed that while most activities on YouTube followed a Poisson distribution, many videos exhibited bursty behaviour with power-law decaying patterns. Matsubara et al. [43] further expanded on these findings by developing a model that explained the rise and fall patterns within information cascades, with the fall phase exhibiting a power-law distribution. Our empirical work [44] shows that a power-law triggering kernel represents user activity well and provides predictive performances.

Temporal point processes have emerged as the leading methodology for modelling user activity on social media platforms. Several recent reviews have provided comprehensive insights into different aspects of temporal point processes. These include: (1) a general review by Lima [45], (2) a review specifically focusing on finance by Hawkes [46], (3) a review on spatio-temporal processes by Reinhart [47], and (4) a review on neural temporal point processes with applications in social media by Shchur et al. [48]. These reviews offer valuable perspectives on the subject matter, covering various topics and highlighting the significance of temporal point processes in various domains.

3.3 Dataset

To showcase the applicability of our methodology, we compiled a Twitter dataset comprising tweets emitted between 1 July and 31 December 2022, relating to two topics often linked to illegal content, as identified by prior literature and news media. The two topics are defined by the hashtags (1) #climatescam (479,051 posts) – controversial opinions regarding climate change [49]; (2) #americafirst or #americansfirst (278,899 posts) – debates over key U.S. political topics such as immigration and foreign policies [50].

3.4 Method

We estimate the relationship between moderation efforts and content-sharing dynamics using a well-established point process framework [48] (see SI for a review).

3.4.1 Model

Events in social media are commonly modelled using self-exciting point processes, also referred to as Hawkes processes [51]. A conditional intensity function $\lambda(t|\mathcal{H}_t)$ defined as

$$\lambda(t|\mathcal{H}_t) = \mu + \sum_{i:t_i<t} \phi(t - t_i; m_i), \quad t \geq 0, \tag{3.1}$$

captures how historical events cause the generation of new events. In this work, events are social media postings – both illegal and non-illegal, which are assumed to share the diffusion dynamics within a topic. The background rate μ explains exogenous effects, such as users spontaneously posting new information; the kernel $\phi(\cdot)$ controls the endogenous dynamics of how users reshare and retweet the content they see. The history $\mathcal{H}_t$ stores the event times t_i when an action was performed and the follower count of the user m_i as $\{(t_i, m_i)\}_{i=1}^{N(t)}$, where $N(t)$ is the total number of events. We model contagion using a power-law function

$$\phi(\tau; m) = \kappa\, m^{\beta} (\tau + c)^{-(1+\gamma)}, \quad \tau \geq 0, \tag{3.2}$$

as the events' influence is long-lasting and has been shown to be heavy-tailed [38]. The memory parameter γ captures the speed at which the content is forgotten. A higher γ indicates the content's importance is fast decreasing. The scalar κ describes the quality of the post. The shift parameter c captures the waiting time before users interact with the post. The exponent β warps the user follower count m. The follower count is known to follow power-law distribution $P(m) = (\alpha - 1)\, m^{-\alpha}$ with parameter α = 2.016 [52]. By integrating Eq. (3.2) over time and social influence, we obtain the potential harm (also known as the *branching coefficient*) as

$$n^* = \int_1^{\infty} \int_0^{\infty} P(m)\, \phi(\tau)\, d\tau\, dm = \kappa \frac{\alpha - 1}{\alpha - \beta - 1} \frac{1}{\gamma\, c^{\gamma}}, \tag{3.3}$$

where $\alpha < \beta - 1$ and $\gamma > 0$. Social media content is described by a subcritical regime, meaning $n^* < 1$, as retweet cascades vanish in the long term. The expected number of events for a subcritical Hawkes process over a long observation horizon T is

$$\mathbb{E}[N(T)] = \mathbb{E}\left[\int_0^T \lambda(t|\mathcal{H}_t)\, dt\right] \approx \frac{\mu T}{1 - n^*}. \tag{3.4}$$

$$\tau_{1/2} = c\,(2^{\frac{1}{\gamma}} - 1)$$

3.4.2 Content Half-Life

The time required by a post to generate 50% of its expected direct offspring, $\tau_{1/2}$, is determined as

$$\int_0^{\tau_{1/2}} \phi(z; m)\, dz = \frac{1}{2} \int_0^{\infty} \phi(z; m)\, dz. \tag{3.5}$$

We substitute Eq. (3.2) in Eq. (3.5), and obtain $\tau_{1/2} = c\,(2^{\frac{1}{\gamma}} - 1)$

3.4.3 Content Removal

Here, we outline the connection between the moderation time Δ and the harm reduction χ. When a post is moderated at deletion time Δ, its direct offspring generation rate drops to zero (see bottom panel of Fig. 3.1). We define the moderated kernel $\phi_\Delta(\tau; m)$:

$$\phi_\Delta(\tau; m) = \begin{cases} \kappa\, m^\beta (\tau + c)^{-(1+\gamma)}, & 0 < \tau \leq \Delta, \\ 0, & \tau > \Delta. \end{cases} \tag{3.6}$$

Hence, we formally define the actual harm as the expected number of direct offspring that a illegal post generates prior to its moderation at time Δ. We compute the actual harm similarly to Eq. (3.3), by replacing $\phi(\tau; m)$ with $\phi_\Delta(\tau; m)$ (cf. Eq. (3.6)):

$$n^*_\Delta = \frac{\kappa}{\gamma} \frac{\alpha - 1}{\alpha - \beta - 1} \left(\frac{1}{c^\gamma} - \frac{1}{(c + \Delta)^\gamma} \right). \tag{3.7}$$

Next, we compute the harm reduction as the percentage of avoided illegal offspring through the branching process. Assuming only one post in the expected event stream (cf. Eq. (3.4)) is illegal, this post will generate many offspring. Not all of these are illegal, as users might try to debunk or respond to the harm. However, we assume all generated illegal posts are moderated in their turn through the DSA mechanism at time Δ. The resulting harm reduction is thus given by

$$\chi = 1 - \frac{\frac{1}{1 - n^*_\Delta}}{\frac{1}{1 - n^*}} = \frac{n^* - n^*_\Delta}{1 - n^*_\Delta}. \tag{3.8}$$

Through the substitution of Eq. (3.3) and Eq. (3.7) into Eq. (3.8), we can reframe the equation to represent Δ as a function of χ,

$$\Delta = \max \left\{ 0, \left(\frac{1}{n^*\, c^\gamma} \frac{\chi (1 - n^*)}{1 - \chi} \right)^{-\frac{1}{\gamma}} - c \right\}. \tag{3.9}$$

Finally, we compute the colourmaps in Fig. 3.2 using Eq. (3.8) and Eq. (3.9), and we project the parameters of the real-world datasets obtained using the estimation procedure (see Statistical Inference).

3.4.4 Statistical Inference

Given observed (real-world) data, the parameters of the Hawkes process specified in Eq. (3.1) – i.e., κ, β and γ – are identified via maximum-likelihood estimation (MLE). Difficulties in parameter estimation arise from extremely flat log-likelihood curves near the optimum and the non-convex objective function (cf. Eq. (3.2)) [53]. To avoid these problems, we perform MLE with multiple starting points in parallel and select the fitted parameters resulting in the highest log-likelihood (see next paragraph). As we can separate original posts from retweets, we estimate the exogenous effect μ from empirical observations and set the shift parameter c to thirty seconds. In Fig. 3.2, the resulting estimates are depicted as point clouds with kernel density estimate (KDE) plots. The objective of maximum-likelihood estimation (MLE) for self-exciting point processes is to maximize the log-likelihood function, given a history of events $\mathcal{H}_t$:

$$\log \mathcal{L}(\theta) = \sum_{i=1}^{k} \log \left(\lambda(t | \mathcal{H}_t) \right) - \int_0^T \lambda(s | \mathcal{H}_s)\, ds,$$

where θ represents a parameter vector, and T the observation horizon. For the proposed model in Eq. (3.2) of the manuscript, considering the parameter vector $\theta = \{\kappa, \beta, \gamma\}$, we obtain

$$\log \mathcal{L}(\theta) = \sum_{i=1}^{k} \log \left(\mu + \sum_{j=1}^{i-1} \phi(t_i - t_j; m_j) \right) - \mu T - \frac{\kappa}{\gamma} \sum_{i=1}^{k} (m_i)^{\beta} \left(c^{-\gamma} - (T - t_i + c)^{-\gamma} \right).$$

Multiple approaches are available to enhance the model and improve its overall fit. However, it is important to note that these options must be tested on specific data and may not necessarily generalize across different discussion topics. One potential approach to consider is incorporating the shift parameter c as part of the parameter vector. This adjustment has shown promise in providing superior fits, particularly when the data are characterized by low user activity. By including c as an adjustable parameter, the model can more accurately capture the data dynamics. Another alternative is to model the background rate as a non-constant factor. Instead of assuming a constant background rate, one can employ a piece-wise function $\mu(t)$ that considers latent information, such as Google trends data, or any other form to capture exogenous behaviour in a more detailed and granular manner. Both approaches offer potential avenues for improving the model's fit to the data. However, it is essential to conduct further exploration and investigation to determine their effectiveness in capturing the complexities of the specific domain under investigation.

3.4.5 Predictive power of temporal point processes

As generative processes, Hawkes point process-based tools can explain the observed past data and predict likely future outcomes. This makes this approach ideally suited for describing the spread of illegal content and predicting its likely spread if not moderated. We present next some literature on which we based this claim. Hawkes processes have found primary application in describing various phenomena in the natural sciences, such as earthquakes [54], the spread of viral diseases like dengue fever [55] and COVID-19 [56], as well as information contagion [57]. In recent years, these processes have also emerged as powerful tools for prediction, capitalizing on the self-exciting property of the kernel. Notably, Mohler et al. [58] employed Hawkes processes to model and forecast the likelihood of crimes. In social media, several Hawkes processes have emerged as state-of-the-art solutions for predicting popularity [59, 60, 61]. Moreover, most current neural temporal point processes for social media applications have been designed for prediction tasks or causal inference from large networks, where latent information renders statistical methods computationally infeasible within reasonable time frames. Another noteworthy application of Hawkes processes is their use in high-frequency trading to model order-book dynamics [62, 63].

"On-the-fly" (real-time) parameter estimates

The parameter fitting procedure we deploy in this work implements the current state-of-the-art in real-time estimation. Next, we describe the approach as used in prior work and our adaptations. The concept of achieving "on-the-fly" parameter estimation, as described in [64], is currently applied in real-world finance applications. This estimation approach can be seen as a first-in, first-out (FIFO) queue of Hawkes events (cf. [65]), defined by the equation:

$$Q(t) = N(t) - N(t - \delta).$$

Here, $N(t)$ represents the counting process, and t denotes the current time. The time interval δ signifies the duration for analyzing content dynamics. One drawback of this formulation is that there is a possibility of having an insufficient number of arrivals within the interval δ. This arrives mainly due to the variability of

the event arrival rate. However, due to the rapidly changing landscape, selecting an optimal fixed horizon δ presents challenges. For instance, [64] explored the optimal horizon δ for describing the endogeneity and exogeneity of cryptocurrency price changes. Therefore, our estimation approach is based on a slightly different queue formulation:

$$Q(t) = N(t) - N(\phi(t)).$$

Here, $\phi(t) : t \to t_{n-k}$, where n represents the last event and κ is the number of events included in the estimation. Designing the system in this way ensures less biased estimates by guaranteeing a sufficient event history.

For achieving unbiased estimates in univariate Hawkes processes, recent research has highlighted the expectation maximization (EM) method as the most powerful approach, as it is more robust to issues such as flat log-likelihoods by leveraging the branching structure of the process [66]. However, this method requires significantly longer computation time due to constructing the branching structure matrix at each iteration. Therefore, a practical approach is to parallelize maximum-likelihood estimation (MLE) and select the process parameters that yield the highest log-likelihood.

To our knowledge, neural temporal point processes for univariate processes do not offer faster computation times. The time complexity for processes with exponential kernels is $O(k)$. The challenge lies in accelerating the computation for power-law processes that do not satisfy the Markov property. This issue has been successfully addressed by Bochud and Challet [67, who approximated the power-law distribution as a sum of weighted exponentials. This approach has been applied to model power-law inter-event time distributions in Hawkes processes that capture trading activity [68, 69, 64]. In our current model, we did not employ this approximation, as it might confuse the reader regarding the technical details and reduce the interpretability of the direct meaning of the power-law decay.

3.4.6 Absolute vs. Relative Harm Reduction

To represent the absolute harm reduction, Eq. (3.8) of the manuscript can be modified as follows:

$$X = \frac{1}{1-n^*} - \frac{1}{1-n^*_\Delta}, \tag{3.10}$$

where X denotes the absolute harm reduction. From Eq. (3.10) we observe that the absolute harm reduction increases with the potential harm n^*. The reason is the following: Content with high n^* tends to generate many direct offspring, which generate their offspring through the branching process. Even for a low half-life time, as for Twitter, this process takes time as multiple generations need to be generated. The branching process is stopped when the moderation is performed (even with Δ = 24 hours). It follows that the higher n^*, the higher the potential number of generations and offspring that are prevented by moderation. As a result, both the absolute and relative harm reduction increase with the potential harm n^*.

3.5 Results

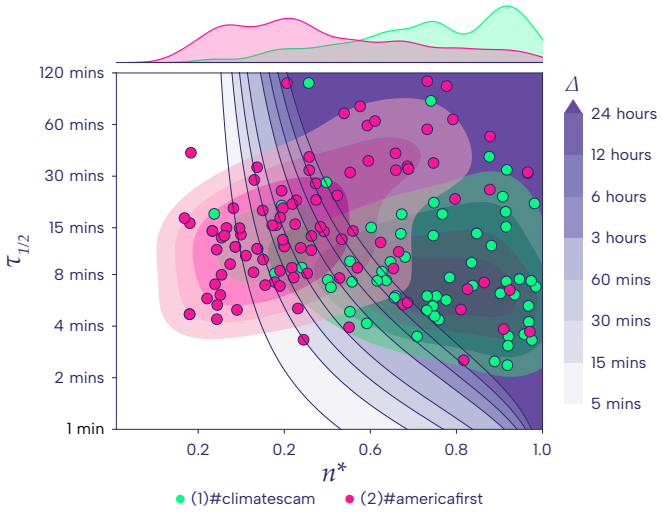
We analyze the likely effectiveness of the EU-regulated moderation as the interplay between the reaction time Δ and the harm reduction χ (see Table 3.1). This interplay is modulated by the online content's potential harm n^* and half-life $\tau_{1/2}$. The colourmap in Fig. 3.2 (a)explores the question **what is the maximum reaction**

time required to achieve a given harm reduction? (here $\chi = 20\%$) On the colourmap, we project real-world discussions around two potentially problematic topics: climate change denial and nationalistic political views (see Materials and Methods). We obtain the positions of the data points by estimating the process dynamics over rolling time windows. The mean content half-life for the two topics is 25.8 mins. The centroids of the discussion dynamics are at ($\tilde{n}^* = 0.75$, $\tilde{\tau}_{1/2} = 7.48$ mins)[1]for #climatescam and ($\tilde{n}^* = 0.44$, $\tilde{\tau}_{1/2} = 13.97$ mins) for #americafirst. This indicates that due to the significantly higher virality for #climatescam, more time ($\Delta >$ 24 hours) is available for content moderation compared to the nationalism topic (Δ = 2.22 hours). Fig. 3.2 (b) explores the question **what is the expected harm reduction when content moderation is performed within 24 hours?** (as currently stipulated by the EU regulation) The harm reduction at the centroids is $\chi = 29.18\%$ and $chi = 13.29\%$ for #climatescam and #americafirst, respectively.

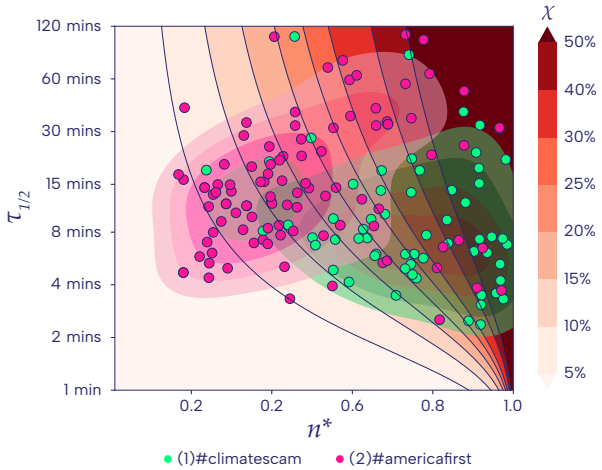
3.6 Discussion

Our work introduces two measures for the effectiveness of DSA moderation: the potential harm n^* and the content half-life $\tau_{1/2}$. For content with known n^* and $\tau_{1/2}$, we can determine the relation between the reaction time Δ and the harm reduction χ.

Figure 3.2: Visual representation of the dependency of the reaction time Δ (a) and harm reduction χ (b) with the potential harm n^* (x-axis) and content half-live $\tau_{1/2}$ (y-axis). Both Δ and χ increase for longer half-lives and higher virality. Real-world potentially problematic content exhibits widely highly variable dynamics. For **#climatescam**, we can achieve harm reduction of [15%, 50%] for Δ = 24 hours.



(a) Reaction time Δ to achieve harm reduction of χ = 20%.



(b) Harm reduction χ when content is removed within Δ = 24 hours.

We make three observations. First, Fig. 3.2 (a) shows the reaction time Δ increases with both half-life $\tau_{1/2}$ and the potential harm n^*. While the former is intuitive, the latter is significant as it indicates that the DSA-legislated moderation can be effective even on Twitter, the platform with the shortest half-life. For example, most of the discussions on #climatescam have high reaction times ($\Delta >$ 24 hours). Second, Fig. 3.2 (b) shows that the harm reduction χ increases with the potential harm n^*. This somewhat counterintuitive result arises from the non-linear interactions between χ, n^* and $\tau_{1/2}$ (see Materials and Methods). It is significant as it indicates that DSA moderation can effectively stop the most illegal content. Third, despite a significant difference in the distribution of potential harm χ for the two topics (see the density marginals on top of Fig.

1 Let $\tilde{x}$ denote the median of x.

3.2 (a)), we see that the most illegal content can emerge in both topics. Therefore, we cannot select a single topic for moderation. Our approach can be used as a strategy to direct the manual flagging efforts towards the most illegal content ($n^* > 0.8$) that can be effectively moderated ($\chi > 50\%$). This would increase the overall effectiveness of the moderation and the flagger's workload.

The major social media platforms employ large content moderation teams, estimated to 15,000 (Facebook), 10,000 (YouTube) and 1,500 (Twitter); each moderator addresses between 600 and 800 claims daily [70]. While there are questions concerning the suitability of the moderators in terms of social context awareness, native language and moderation guidelines, the raw numbers indicate that the platforms already employ the required manpower to implement the new legislation. With the DSA, the European Union seeks to make this process uniform across platforms, transparent and regulated. The keys to its success seem to be appointing trusted flaggers, developing an effective tool for reporting illegal content across platforms, and correctly timing the reaction time for moderation. This chapter provides a novel framework for policymakers to draft mechanisms for content moderation by indicating where to focus human fact-checking efforts and how quickly to react.

4. Synthetic Testbed: Stemming Far-Right Opinion Spread Using Positive Interventions

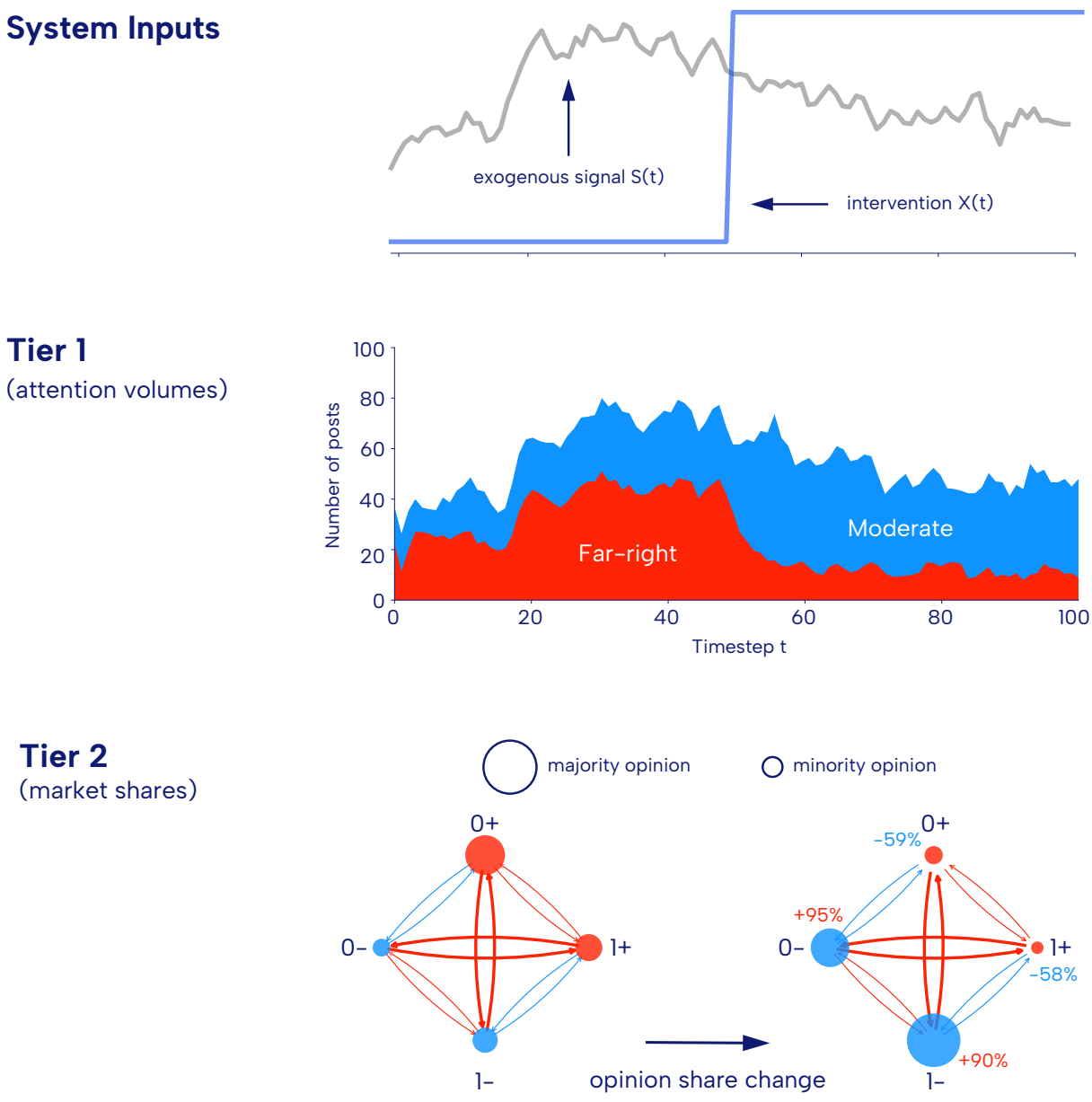
Online extremism has severe societal consequences, including normalizing hate speech, user radicalization, and increased social divisions. Various mitigation strategies have been explored to address these consequences. One such strategy uses positive interventions: controlled signals that add attention to the opinion ecosystem to boost certain opinions. In this chapter, we evaluate the effectiveness of positive interventions. We introduce the Opinion Market Model (OMM), a two-tier online opinion ecosystem model that considers both inter-opinion interactions and the role of positive interventions. The size of the opinion attention market is modelled in the first tier using the multivariate discrete-time Hawkes process; in the second tier, opinions cooperate and compete for market share, given limited attention using the market share attraction model. We demonstrate the convergence of our proposed estimation scheme on a synthetic dataset. Next, we test OMM on two learning tasks, applying to two real-world datasets to predict attention market shares and uncover latent relationships between online items. The first dataset comprises Facebook and Twitter discussions containing moderate and far-right opinions about bushfires and climate change. The second dataset captures popular VEVO artists' YouTube and Twitter attention volumes. OMM outperforms the state-of-the-art predictive models on both datasets and captures latent cooperation-competition relations. We uncover (1) self and cross-reinforcement between far-right and moderate opinions on the bushfires and (2) pairwise artist relations that correlate with real-world interactions such as collaborations and long-lasting feuds. Lastly, we use OMM as a testbed for positive interventions and show how media coverage modulates the spread of far-right opinions.

4.1 Introduction

Online social media platforms are fertile grounds for deliberation and opinion formation [71, 72]. Opinions thrive in the online opinion ecosystem, interconnected online social platforms where they interact – cooperating or competing for the finite public attention [73]. We delineate two types of interventions to mitigate the spread of extremist views. Negative interventions aim to subtract attention from the opinion ecosystem by placing fact-check warnings on postings [74], shadowbanning [75] or outright banning extremist social media groups and accounts [76]. While negative interventions are effective [77], they are available solely to the social media platforms that tend to use them sparingly [78].

Positive interventions [79], such as misinformation debunking [80, 81] and increasing media coverage [82], mitigate extremist views by adding attention to the online opinion ecosystem through informing the public, redistributing attention away from extremist, and toward moderate views. Such interventions are typically in the hands of government and media agencies [83]. Testing the viability of positive interventions requires capturing reaactions to interventions and inter-opinion interactions.

Figure 4.1: We illustrate how the positive intervention $X(t)$ (defined in Eq. (4.9)) suppresses far-right opinions on a simulated toy opinion ecosystem with two far-right (0+, 1+) and two moderate (0−, 1−) opinions. For instance, 0+ and 1+ can represent the opinions "the Greens policies caused the Australian bushfires" and "mainstream media cannot be trusted," respectively; 0− and 1− can be obtained as their negations. Top row: the exogenous signal $S(t)$ (defined in Eq. (4.5)) and the intervention $X(t)$. Middle row: total daily opinion market size quantified by our model's first tier, split into far-right (+) and moderate (−) opinion volumes. Bottom row: market shares and the interactions between the four opinions estimated by our model's second tier. Nodes are opinions; their sizes indicate market share; edges represent exciting (red) and inhibiting (blue) relations. $X(t)$ suppresses far-right opinions for $t > 50$. Shown are average market shares before (left) and after (right) $t = 50$.



This work develops a model for the dynamics of the opinion ecosystem and a testbed for evaluating positive interventions. We focus on two open questions. The first question explores the analogy between opinions and economic goods. In a competitive economic market of limited resources, coexisting goods can interact in one of two ways: either they compete for market share (substitute brands, like Pepsi and Coke) or reinforce each other (complementary items, like bread and butter). We argue that opinions in the online ecosystem behave similarly, allowing us to leverage market share modelling tools [84]. The first research question is: **Can we model the online opinion ecosystem as an environment where opinions cooperate or compete for**

market share? We propose the Opinion Market Model (OMM), a two-tier model to address this question. Fig. 4.1 showcases a simple opinion ecosystem under intervention, with two opinions (denoted 0 and 1) on a single social media platform. Each opinion has two polarities: far-right supporters (+) and moderate debunkers (-). Exogenous signals (shown in grey in the top panel of Fig. 4.1) and interventions (shown in yellow) modulate the dynamics of the opinions' sizes. Exogenous signals are naturally occurring events like bushfires, floods, or political speeches. Interventions (like increased media coverage) are designed to add attention to the opinion ecosystem, increasing the market share of certain opinions while suppressing others. The first tier of OMM (middle row in Fig. 4.1) uses a discrete-time Hawkes process to estimate the size of the opinion attention market – that is, the daily number of postings featuring opinions. The Hawkes process has been widely used to model online attention [61, 85] due to its ability to account for exogenous factors and the endogenous "word-of-mouth" through its self- and cross-exciting property. The second tier of OMM (bottom row in Fig. 4.1) leverages a market share attraction model to capture opinion interactions – we assume that online opinions compete for the users' limited online attention [86, 87]. For the example in Fig. 4.1, opinions 0- and 1+ have a strong reinforcing relation (shown as red arrows), while 1- and 1+ have a weak competing relation (blue arrows).

We test OMM on two real-world datasets. The first contains Facebook and Twitter discussions expressing moderate and far-right opinions on bushfires and climate change [88]. The second captures the YouTube views and Twitter mentions for the most popular VEVO artists' songs in 2017 [73]. We evaluate OMM on two tasks: predicting attention market share and exposing relationships between online items. For the predictive task, OMM outperforms the current state of the art in market share modelling (Correlated Cascades [85] and Competing Products [89]) and predictive baselines on both datasets. For the second task, we leverage the OMM to expose the relations between opinions on the two platforms. For the bushfire case study, no significant interactions occurred on Facebook, as postings were collected from far-right public groups with limited interaction with the opposing side. On Twitter, we observe self-reinforcement behaviour of both far-right and moderate opinions, probably due to the echo chamber effect [90] – reinforcing one's beliefs due to repeated interactions with users sharing similar ideologies on social platforms. Surprisingly, we notice that opposing views reinforce each other, probably due to the deliberative nature of Twitter, where far-right sympathizers and opponents oppose each other. For the VEVO artists case study, we uncover nontrivial pairwise interactions of music artists correlating with real-world relationships – such as Ariana Grande's and Calvin Harris' reinforcement relationship due to their collaboration "Heatstroke" and Taylor Swift's and Justin Bieber's inhibiting relationship.

Our second research question is: **Can we test the sensitivity of the opinion ecosystem to positive interventions?** OMM accounts for positive interventions – controlled external signals to boost certain opinions. In Fig. 4.1 an intervention is performed for $t > 50$, which suppresses the far-right opinions (+), leading to the shrinking of their market share. We use OMM for two tasks: first, to estimate whether interventions effectively shape the opinion ecosystem and, second, to construct what-if scenarios as synthetic testbeds for future interventions. For the bushfire case study, we test whether news coverage from reputable and controversial media outlets suppresses or aids the spread of far-right opinions. We fit OMM twice: with and without media coverage. We find a better fit with the intervention, suggesting that media coverage actively shapes the opinion ecosystem. We perform synthetic what-if experiments: we vary the level of media coverage, simulate the system and observe the effect on opinion market shares. On Facebook, reputable media coverage reduces the prevalence of far-right opinions. On Twitter, both reputable and controversial media coverage suppress far-right opinions. However, for some opinions (like "Mainstream media cannot be trusted"), reputable news backfires increasing far-right opinions share.

The main contributions of the work are as follows:

1. A novel two-tier model of the opinion ecosystem that allows studying opinion interactions through an economics-based cooperation-competition lens. We introduce simulation and estimation algorithms and study the convergence with synthetic tests.
2. A synthetic testbed to uncover interactions across sympathizers and opponents of far-right opinions and likely effects of positive interventions via media coverage.
3. A curated dataset of Twitter and Facebook discussions on bushfires/climate change.

Related Work. We focus the discussion of related work on models for cooperative-competitive interaction in a set of co-diffusing online items. These models need to be both predictive and interpretable (usually generative models). We have observed a lack of recent research in this area, with few works dating after 2017. Closely related to our proposal is the Correlated Cascades (CC) model [85], a variant of the multivariate Hawkes process to model product adoption across a set of competing products in a social network. It estimates the interaction parameter β, tuning the market cooperation or competition level. A limitation of CC is that all products share a single β value. This simplifies existing asymmetric relationships and assumes that all brands either cooperate or compete. OMM addresses this issue by fully modelling these asymmetric relationships. Another closely related work is the Competing Products (CP) model [89], a multivariate Hawkes model for product adoption/use where the frequency of use is affected by the usage of other products. Limitations of the work are the absence of the assumption of limited attention and the possibility of negative intensities since competitive interactions are modelled as negative parameters. OMM avoids the weaknesses of CP by using a multiplicative model, thereby avoiding negative intensities and defining opinion shares as fractions of the total attention volume. The SLANT model [91] and the follow-up SLANT+ [92] extend the CP model to differentiate between a user's latent and expressed opinion and uses a similar Hawkes process to model the intensity. However, SLANT requires fine-grained network information for training, which is prohibitive considering that online platforms are becoming more stringent with fine-grained data access [93]. On the other hand, OMM requires only opinion counts for training.

Ethics of Opinion Moderation and Broader Perspectives. OMM is intended to model interactions between opinions and be used as a testbed for evaluating positive interventions for opinion moderation. As any tool, OMM is unaware of the intention of its user and, in theory, could be used by oppressive regimes to silence or manipulate the liberal opinions of their opponents [83]. In addition, the fundamental value of freedom of speech for democratic societies implies that non-widely accepted opinions also have the right to be expressed. The scientific literature studies this ethical conundrum in the context of Countering Violent Extremism (CVE) initiatives [94, 83]. When viewing OMM as an AI evaluation tool supporting CVE initiatives [95], these ethical issues can be mitigated using online CVE frameworks in liberal democracies [96]. We argue that the implementing body is responsible for OMM's ethical usage, and CVE regulations should be leveraged to mitigate malicious intent.

Causal Impact. OMM measures the effect of media coverage on the opinion market shares using a generative model to disentangle endogenous and exogenous factors from observational data, similar to [97, 98, 99]. Our model works on aggregate observational data (i.e., opinion counts), and it does not prove the causal impact of media coverage on individual opinion formation (i.e., behaviour change). We would require a pre-test/post-test control group design to achieve true causal links. Previous work [100, 101, 102] provides evidence of

the interventional role of media coverage. In Section 4.8, we explore this further in a what-if experiment to demonstrate how the level of media coverage affects opinion market shares.

4.2 Preliminaries

We introduce two classes of models that form the foundation of our approach: (1) the discrete-time Hawkes process [103], a model of event counts that display self-exciting behaviour, and (2) the market share attraction model [84], a marketing model that uncovers the latent competitive structure of brands and interaction with marketing instruments.

4.2.1 Discrete-time Hawkes Process

The discrete-time Hawkes Process (DTHP) [103] is the discrete-time analogue of the popular self-exciting Hawkes process [104], where instead of modelling the occurrence of events given by $t \in \mathbb{R}^+$, we model the event count $N(t)$ on $[t-1, t)$ for $t \in \mathbb{N}$.

The DTHP is characterized by the conditional intensity function $\lambda(t)$, defined as the expected number of events that occur at time t, conditioned on the history $H_{t-1} = \{N(s) | s < t\}$. For a DTHP, $\lambda(t)$ is given by

$$\lambda(t) = \mathbb{E}[N(t)|H_{t-1}] = \mu + \sum_{s<t} \alpha \cdot f(t-s) \cdot N(s), \tag{4.1}$$

where μ is the baseline count of events, α determines the level of self-excitation and is the expected number of events produced by a single event and $f(t)$ is the triggering kernel, which controls the influence of the past events on the present. We specify $f(t)$ with the geometric probability mass function $f(t) = \theta(1-\theta)^{t-1}, t \in \mathbb{N}$, the discrete-time analogue of the exponential distribution [103]. Given $\lambda(t)$, model specification is completed by specifying a probability mass function for the count $N(t)$. Following [103], we set $N(t) \sim \text{Poi}(\lambda(t))$.

4.2.2 Market Share Attraction Model

In marketing literature, market share attraction models (MSAMs) [84] model the competitive structure of a set of M brands in a product category, predict their market shares, and evaluate how a set of marketing instruments affects resulting market shares. MSAMs assume that the market share s_i of brand $i \in \{1 \dots M\}$ is proportional to consumers' attraction $\mathcal{A}_i$ towards brand i:

$$s_i = \frac{\mathcal{A}_i}{\sum_{j=1}^{M} \mathcal{A}_j} \in [0, 1]. \tag{4.2}$$

$\mathcal{A}_i$ is typically modelled as a parametric function of a set of K marketing instruments $\{X_{ki}\}_{k=1}^{K} \in \mathbb{R}^K$, where X_{ki} gives the value of the k^{th} marketing instrument for brand i. We typically specify $\mathcal{A}_i$ as

$$\mathcal{A}_i = \exp\left(\beta_i + \sum_{k=1}^{K} \sum_{j=1}^{M} \gamma_{kij} X_{kj}\right), \tag{4.3}$$

where β_i measures the inherent attraction of brand i and $\gamma_{kij} \in \mathbb{R}$ measures the effect of the value of the k^{th} marketing instrument for brand j on brand i's attraction. Whether γ_{kij} is positive (negative) is indicative of the excitatory (inhibiting) relationship from brand j to brand i through marketing instrument X_{kj}.

MSAMs are interpreted via the model elasticity $e(s_i, X_{kj})$, the ratio of the percent change in the market share s_i given a percent change in the value of the k^{th} marketing instrument for brand j. For example, an elasticity of $e(s_i, X_{kj}) = 0.1$ means that a 1% increase in X_{kj} corresponds to a 0.1% increase in s_i. That is,

$$e(s_i, X_{kj}) = \frac{\partial s_i / s_i}{\partial X_{kj} / X_{kj}} = \frac{\partial s_i}{\partial X_{kj}} \cdot \frac{X_{kj}}{s_i}. \tag{4.4}$$

The elasticity $e(s_i, X_{kj})$ captures the overall effect of brand j's marketing instrument X_{kj} on brand i's market share s_i: both the direct effect of X_{kj} on s_i, controlled by γ_{kij}, and the indirect effect of X_{kj} on s_i through its effect on the attraction of other brands $\{j \neq i\}$.

4.3 The OMM Model

In this section, we develop a two-tier model of the opinion ecosystem. The first tier models the total size of the opinion attention market on multiple online platforms. The second tier models the market share of opinions on each platform. Next, we introduce a scheme for parameter estimation.

OMM consists of two tiers; the first tier, which we call the opinion volume model, tracks the size of the opinion attention market, while the second tier, the opinion share model, tracks the market shares of the different opinions. Table 4.1 summarises the notation for important variables in the OMM. The full table is available in the online appendix [105].

Opinion Volume Model. Suppose our opinion ecosystem consists of P social media platforms. The opinion volume model tracks the attention volume, i.e. the number of opinionated posts $N^p(t)$, on each platform $p \in \{1, \dots P\}$ and time $t \in \mathbb{N}$. We model $\{N^p(t)\}_p$ as a P–dimensional DTHP (defined analogous to the multivariate Hawkes process [104]) with conditional intensity $\{\lambda^p(t)\}_p$,

$$\lambda^p(t) = \mu^p \cdot S(t) + \sum_{q=1}^{P} \sum_{s<t} \alpha^{pq} \cdot f(t-s) \cdot N^q(s). \tag{4.5}$$

In contrast to Eq. (4.1), we use a time-varying exogenous signal $S(t)$, which accounts for the baseline volume of events of exogenous origin. The signal $S(t)$ accounts for natural tendencies and events (i.e., epidemics, elections) and typically cannot be controlled. We introduce a scaling term μ^p for each platform p such that $\mu^p \cdot S(t)$ represents the exogenous opinion count for platform p on time t.

Since online platforms are not siloed and have significant user overlap, we allow the P platforms to interact via intra- and inter-platform excitation. The parameter $\alpha^{pq} > 0$ sets the level of intra-platform (for $p = q$) and inter-platform (for $p \neq q$) excitation. Lastly, we set $N^p(t) \sim \text{Poi}(\lambda^p(t))$.

Opinion Share Model. With the attention volumes for each platform p estimated in the first tier, the second tier models the market share $s_i^p(t)$, calculated as the fraction of opinionated posts on platform p conveying opinion i. Given the limited attention market size, opinions compete for attention within each platform.

Suppose that there are M different opinion types. We set $N_i^p(t)$ to be the number of opinionated posts conveying opinion i on platform p on time t, and $\lambda_i^p(t)$ to be its conditional intensity. Using the notion of limited attention [85], we relate $\lambda_i^p(t)$ to $\lambda^p(t)$ in Eq. (4.5) by introducing the market share $s_i^p(t) \in [0,1]$ as the fraction of opinion i posts on platform p. That is,

$$\lambda_i^p(t) = \lambda^p(t) \cdot s_i^p(t), \tag{4.6}$$

and $\sum_{i=1}^{M} s_i^p(t) = 1$.

Similar to Eq. (4.2), we model $s_i^p(t)$ with attraction $\mathcal{A}_i^p(t)$,

$$s_i^p(t) = \frac{\mathcal{A}_i^p(t)}{\sum_{j=1}^{M} \mathcal{A}_j^p(t)}. \tag{4.7}$$

Leveraging the MNL form in Eq. (4.3), we define attraction

$$\mathcal{A}_i^p(t) = \exp \mathcal{T}_i^p(t), \tag{4.8}$$

where $\mathcal{T}_i^p(t)$ consists of two parts, accounting for interventions and endogenous dynamics, and is described in detail below,

$$\mathcal{T}_i^p(t) = \underbrace{\sum_{k=1}^{K} \gamma_{ik}^p \cdot \bar{X}_k(t)}_{\text{interventions}} + \underbrace{\sum_{q=1}^{P} \sum_{j=1}^{M} \beta_{ij}^{pq} \cdot \lambda^q(t|j)}_{\text{endogenous}}, \tag{4.9}$$

$$\bar{X}_k(t) = \sum_{s<t} f(t-s) \cdot X_k(s), \text{ and}$$

$$\lambda^p(t|j) = \mu_j^p \cdot S(t) + \sum_{q=1}^{P} \sum_{s<t} \alpha^{pq} \cdot f(t-s) \cdot N_j^q(s), \tag{4.10}$$

where $\mu^p = \sum_{j=1}^{M} \mu_j^p$.

In the first term of Eq. (4.9), we introduce a set of K positive interventions $\{X_k(t)\}_k$ that modify the opinion market shares in the opinion ecosystem. The interventions $\{X_k(t)\}_k$ have a different to $S(t)$ in Eq. (4.5), as the latter modifies the attention market size. Parameter $\gamma_{ik}^p \in \mathbb{R}$ measures the direct effect of the k^{th} intervention on the market share of opinion i on platform p. If γ_{ik}^p is positive (negative), then $X_k(t)$ reinforces (inhibits) opinion i on platform p.

In the second term of Eq. (4.9) we model the contribution of endogenous dynamics on the attraction of opinion i. To represent the prevalence of opinion j on platform q, we make use of the conditional intensity $\lambda^p(t|j)$ in Eq. (4.10), which models the dynamics of opinion j independent of other opinions. Parameter $\beta_{ij}^{pq} \in \mathbb{R}$ captures the direct effect that opinion j on platform q has on the market share of opinion i on platform p. Similar to γ_{ik}^p, we allow β_{ij}^{pq} to be positive (negative), representing a reinforcing (inhibiting) relationship from opinion j to i on platform q and p, respectively.

Table 4.1: Summary of important quantities and notations.

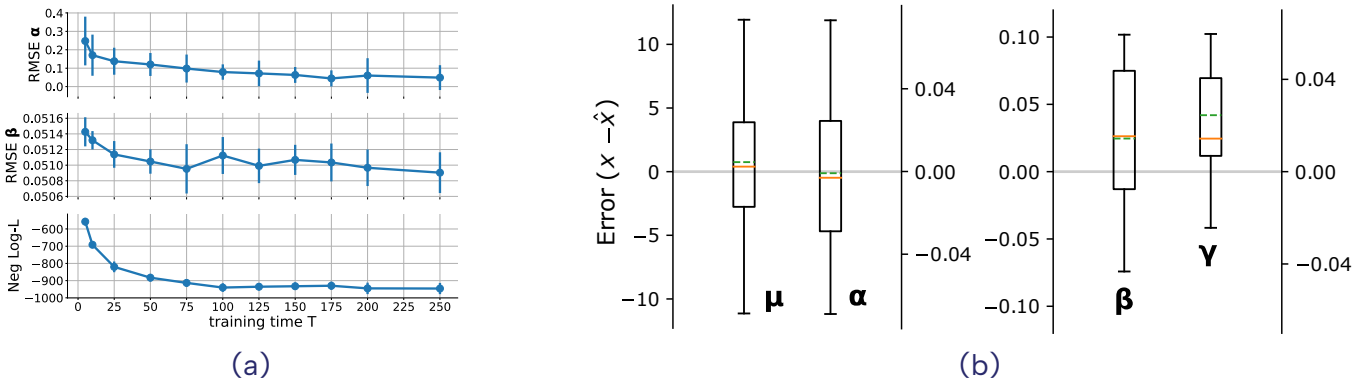
Notation	Interpretation
P	number of social media platforms
M	number of opinion types
K	number of positive interventions
T	terminal time
Variables	
$S(t)$	input signal, volume of exogenous events
$X_k(t)$	input signal, k^{th} positive intervention
$s_i^p(t)$	market share of opinion i on platform p at time t
$\lambda_i^p(t)$	conditional intensity of opinion i
$N_i^p(t)$	#posts with opinion i on platform p at time t
$e(s_i^p(t), \cdot)$	opinion share model elasticity
Data	
$n_t^p / n_{i,t}^p$	#posts on platform p at time t / with opinion i
$s_{i,t}^p$	fraction of posts on platform p with opinion i at time t
Parameters	
μ_j^p	exogenous scaling term for opinion j on platform p
α^{pq}	excitation parameter for intra-platform ($p = q$) and inter-platform (for $p \neq q$) dynamics
θ	memory parameter, describing how fast an event is forgotten, $\theta \in [0, 1]$
γ_{ik}^p	direct effect of the k^{th} intervention on share of opinion i on platform p
β_{ij}^{pq}	direct effect that opinion j on platform q has on share of opinion i on platform p.

Estimation. Over the observation period $t \in \{1, \ldots, T\}$, assume that we observe the exogenous signal $S(t)$ the K interventions $\{X_k(t)\}_k$, and the number $n_{i,t}^p$ of posts conveying opinion i on platform p for each i and p. Our goal is to estimate the parameter set $\mathbf{\Theta} = \{\mu_j^p, \alpha^{pq}, \theta, \gamma_{ik}^p, \beta_{ij}^{pq}\}$.

The structure of our two-tier model allows us to cast parameter estimation as a two-tier optimization problem. Let $\mathbf{\Theta}_1 = \{\mu^p, \alpha^{pq}, \theta\}$. The key observation here is that the first-tier parameter set $\mathbf{\Theta}_1$ can be estimated using

only the opinion volume model in Eq. (4.5), independent of the opinion share model in Eq. (4.9). By optimizing the likelihood $\mathcal{L}_1(\mathbf{\Theta}_1|\{n_t^p\}_{p,t})$ of the platform-level volumes $\{n_t^p\}_{p,t}$, we can obtain an estimate $\hat{\mathbf{\Theta}}_1$ of $\mathbf{\Theta}_1$.

Figure 4.2: Parameter recovery results on synthetic data. In (a), we show the convergence of the RMSE of the α and β estimates and the negative log-likelihood as we increase the training time T. In (b), we show the difference between our estimates for $\{\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}\}$ and the true values. Dashed **green** lines and **orange** lines are the mean and median values, respectively.



(a) (b)

The second-tier parameter set $\mathbf{\Theta}_2 = \{\mu_j^p, \gamma_{ik}^p, \beta_{ij}^{pq}\}$ can be obtained by optimizing the likelihood $\mathcal{L}_2(\mathbf{\Theta}_2|\hat{\mathbf{\Theta}}_1, \{n_{i,t}^p\}_{i,p,t})$ of the opinion volumes $\{n_{i,t}^p\}_{i,p,t}$, conditioned on our estimate of the first-tier parameters $\hat{\mathbf{\Theta}}_1$. Our full estimated parameter set is given by $\hat{\mathbf{\Theta}} = \hat{\mathbf{\Theta}}_1 \cup \hat{\mathbf{\Theta}}_2$. The technical details of the estimation and the derivation of the likelihoods $\mathcal{L}_1(\cdot)$ and $\mathcal{L}_2(\cdot)$ and gradients $\partial_{\mathbf{\Theta}_1}\mathcal{L}_1(\cdot)$ and $\partial_{\mathbf{\Theta}_2}\mathcal{L}_2(\cdot)$ are available in the online appendix [105].

Simulation. Suppose we are given the opinion volume $n_{i,0}^p$ at time $t = 0$ for each platform p and opinion i, such that $n_t^p = \sum_i n_{i,t}^p$. A sample of $n_{i,t}^p$ from OMM can be obtained by calculating the conditional intensity $\lambda_i^p(t)$ using Eq. (4.6), and then sampling $n_{i,t}^p$ from $Poi(\lambda_i^p(t))$. We obtain $\{n_{i,t}^p\}_{i,p,t}$ by repeating these steps over $\{1, \ldots, T\}$.

Numerical Considerations. To improve model fit in our real-world case studies, we implement three augmentations to the model and estimation method, outlined below and fully detailed in the online appendix [105]. First, we modify the attraction $\mathcal{A}_i^p(t)$ in Eq. (4.7) to prevent numerical overflow/underflow. Second, we add a regularization term in the second-tier optimization problem in Section 4.3 to impose structural constraints on $\{\hat{\gamma}_{ik}^p\}$ and improve estimation. Third, we apply log-scaling on $\lambda^q(t|j)$ and standardize both $\lambda^q(t|j)$ and $\bar{X}_k(s)$ in Eq. (4.9) to solve scaling issues.

Stability Assumption. We implicitly assume that the opinion attention market is stable over the timeframe of the analysis, in the sense that the parameters $\mathbf{\Theta}$ governing the behaviour of the process stay constant within the timeframe. In situations where this assumption is not expected to hold (e.g. extreme events) and parameters change within the timeframe, a change-point model extension [106, 103] of the OMM is necessitated.

4.4 Learning with Synthetic Data

In this section, we consider the parameter estimation task with synthetic data. First, we discuss our experimental setup and the synthetic dataset. Next, we show that parameter recovery error decreases and stabilizes as we increase the training time T and the number of samples $n_{samples}$.

Experimental Setup. We set $P = M = K = 2$. We set $\mu_1^1, \mu_2^1, \mu_1^2, \mu_2^2] = [15, 5, 5, 20]$, and $\theta = 0.5$ and draw $\alpha^{pq} \sim \text{Unif}(0, 0.5)$, $\beta_{ij}^{pq} \sim \text{Unif}(0, 0.1)$ and $\gamma_{ik}^{p} \sim \text{Unif}(0, 0.1)$. The exogenous signals are $S(t) = 1$, $X_1(t) = 5\sin(0.1x) + 5$, and $X_2(t) = 10\sin(0.05x + 1.25) + 10$.

We construct our synthetic dataset using the simulation algorithm in Section 4.3 to get 400 samples of opinion volumes $\left\{n_{i,t}^{p}\right\}_{i,p,t}$ for $t \in \{1, \ldots, T = 300\}$. We implement joint fitting [107]: we partition the 400 samples into 20 groups of nsamples = 20 samples each. The likelihoods $\mathcal{L}_1(\mathbf{\Theta}_1)$ and $\mathcal{L}_2(\mathbf{\Theta}_2|\mathbf{\Theta}_1)$ of each group are maximised to get an estimate $\hat{\mathbf{\Theta}}$, yielding 20 sets of parameter estimates.

Model Evaluation. To study the convergence of our learning algorithm, we compute the root-mean-squared error (RMSE) of our estimated $\hat{\mathbf{\Theta}} = \{\hat{\mu}_j^p, \hat{\alpha}^{pq}, \hat{\theta}, \hat{\gamma}_{ik}^p, \hat{\beta}_{ij}^{pq}\}$ with respect to the true $\mathbf{\Theta}$, following [89]. We report the average RMSE per parameter type, where the average is taken over the components of the matrix or tensor corresponding to the parameter type.

In Fig. 4.2 (a), we see that training on a longer timeframe leads to lower RMSE for $\hat{\alpha}^{pq}$ and $\hat{\beta}_{ij}^{pq}$ and better model fit measured by the likelihood $\mathcal{L}_2$. Results for $\hat{\mu}_j^p$, $\hat{\theta}$ and $\hat{\gamma}_{ik}^p$, and on varying nsamples are in the online appendix [105].

In Eq. (4.2) (b), we plot the difference distribution between our estimates and the true values. We recover first-tier parameters $\{\hat{\mu}_j^p, \hat{\alpha}^{pq}\}$ well, as evidenced by our mean estimates coinciding with the true values. We observe a slight overestimation of $\{\hat{\gamma}_{ik}^p, \hat{\beta}_{ij}^{pq}\}$, given the nonconvexity of $\mathcal{L}_2$ and the high dimensionality of the second-tier parameter set.

4.5 Real-World Datasets

This section introduces two real-world datasets we have curated to evaluate the OMM.

4.5.1 Bushfire Opinions dataset

We construct the Bushfire Opinions dataset, containing 90 days of Twitter and Facebook discussions about bushfires and climate change between November 1, 2019 to January 29, 2020. The Facebook postings are a subset of the SocialSense dataset [88]; we select posts and comments about bushfires and climate change (SocialSense also contains discussions around COVID-19). These were collected using CrowdTangle[1] by crawling public far-right Australian Facebook groups, identified via a digital ethnographic study (see [88] and the online appendix [105] for details). We build the Twitter discussions using the Twitter Academic v2 API; we collect tweets emitted between November 1, 2019 to January 29, 2020 that mention bushfire keywords such as

1 https://www.crowdtangle.com/

bushfire, arson, australiaburns (see the full list in the online appendix [105]). We use the AWS Location Service[1] to geocode users based on their free-text location and description fields and filter only for tweets from Australian users.

Our focus on the 2019-2020 Australian bushfires is motivated by the availability of human-annotated topics, opinions [88] and stance classifiers [108] trained on the same topic and timeframe. We use these classifiers to filter and label our dataset.

Moderate and Far-Right Opinion Labeling. To filter and label relevant Facebook and Twitter postings, we use the textual topic and opinion classifiers developed by Kong et al. [88], with a reported 93% accuracy in classifying Facebook and Twitter posts on bushfires and climate change. We select the following most prevalent six opinions, covering 95% of Twitter and 81% of Facebook postings:

0. Greens policies are the cause of the Australian bushfires.
1. Mainstream media cannot be trusted.
2. Climate change crisis is not real / is a UN hoax.
3. Australian bushfires and climate change are not related.
4. Australian bushfires were caused by random arsonists.
5. Bushfires are a normal summer occurrence in Australia.

Furthermore, we deploy the far-right stance detector introduced by Ram et al. [108] – which leverages a textual homophily measurement to quantify the similarity between Twitter users and known far-right activists. On the **Bushfire Opinions Twitter dataset**, the stance detector achieves a 5-fold CV AUC ROC score of 0.889. An opinion is labelled as **far-right** if the posting agrees with the opinion (denoted as +), or **moderate** if the posting disagrees with the opinion (-). We represent our opinion set as $\{(i-, i+) | i \in \{0, \ldots, 5\}\}$. In summary, we consider $P = 2$ platforms with 74,461 tweets and 7,974 Facebook postings labeled with $M = 12$ stanced opinions. We aggregate posting volumes by the hour, resulting in $T = 2,160$ time points over 90 days from Nov 1, 2019, to Jan 29, 2020.

Exogenous Signal S and Intervention X. The exogenous signal $S(t)$ (Eq. (4.5)) modulates the total size of the attention market in the first tier of OMM. We use the 5-day rolling average of the Google Trends query bushfire+climate change in Australia, normalized to a max value of 1. Google Trends captures the baseline interest on topics [109] and is a proxy for offline events (ex. actual bushfires and government measures) [110].

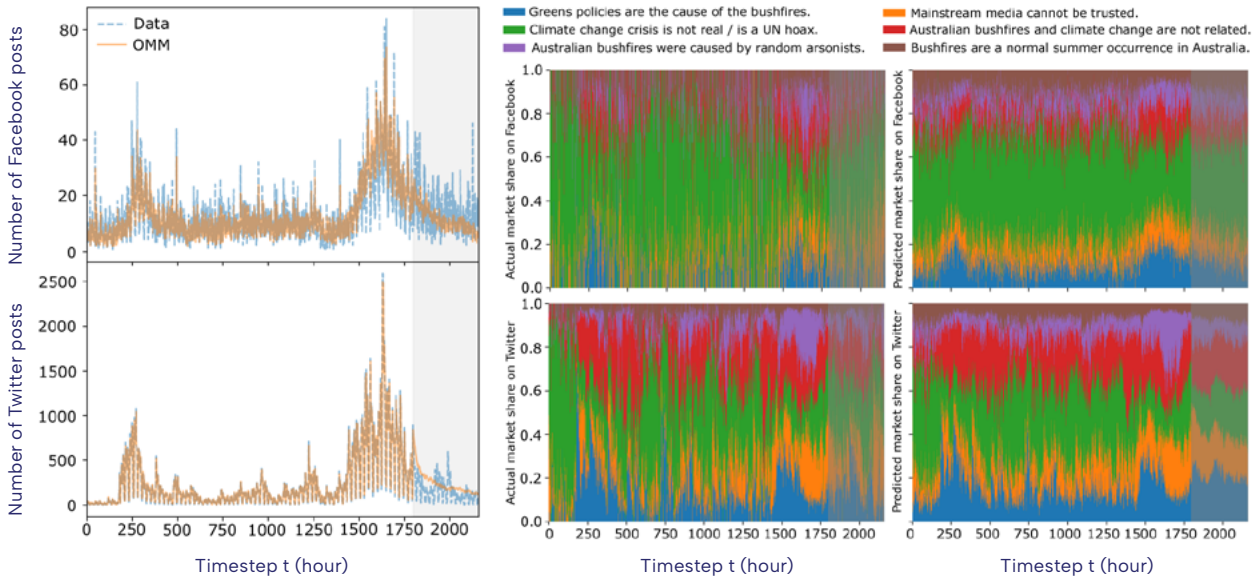
The interventions $\{X_k(t)\}$ modulate the market share of far-right and moderate opinions. Our interventions consist of two sources of news coverage: reputable (R) mainstream Australian publishers (e.g., The Sydney Morning Herald, Canberra Times, Crikey) and controversial (C) international publishers (e.g., Sputnik News, Breitbart, Red State). For each opinion $i \in \{0, \ldots, 5\}$, we consider a pair of interventions $(R_i(t), C_i(t))$, consisting of reputable and controversial daily news volumes discussing opinion i. We assemble the intervention set $\{X_k(t)\}$ $K = 12$ so that the first six interventions correspond to $\{R_0(t), \ldots, R_5(t)\}$ while the last six correspond to $\{C_0(t), \ldots, C_5(t)\}$.

We sourced reputable Australian news publishers from the Reputable News Index (RNIX) [111]. We query Factiva [112] to obtain the daily news volume of these outlets for each of the six opinions using a keyword search. We

1 https://aws.amazon.com/location/

similarly obtain the news volumes from controversial international publishers from NELA-GT-2019 [113] using a keyword search. We subtract the Google Trends signal from the news volumes for each intervention. We compute the standardized form of $X_k(t)$ as $\hat{X}_k(t) = \text{news}_k(t) - \frac{\max_t \text{news}_k(t)}{\max_t S(t)} S(t)$. For brevity, in the bushfire case study, we denote $\hat{X}_k(t)$ as $X_k(t)$ (i.e., always in standardized form). Standardization allows $X_k(t)$ to be interpreted as the extent to which reputable or controversial media over- or under-reports relative to the public's attention.

Figure 4.3: Fitting and predicting with OMM on the Bushfire Opinions dataset. We train OMM on the first 1800 timesteps and predict on timesteps 1801 to 2160 (shaded area). We show results for Facebook (top row) and Twitter (bottom row). (a) Actual (dashed **blue** lines) vs. fitted/predicted (**orange** lines) volumes; (b) Actual (left panels) and fitted during training and predicted during testing (right panels) opinion market shares on Facebook and Twitter. We aggregate the far-right and moderate opinions.



4.5.2 Dataset construction

The **Bushfire Opinions dataset** consists of Twitter posts and Facebook posts & comments from Australian user accounts and pages expressing problematic opinions on climate change and the 2019-2020 Australian bushfire season during the 90-day period of November 1, 2019 to January 29, 2020.

The Bushfire Opinions dataset derives from the **SocialSense dataset** introduced in [88], which consists of user posts and comments from three major online social media platforms: Facebook, Twitter and Youtube. Postings included in the SocialSense were on two general topics – first, the Australian bushfires and climate change, and second, Covid-19 and vaccination – and expressed problematic opinions. In this work, we focus on Facebook/ Twitter and the Australian bushfires/ climate change topic. Postings were collected using Crowdtangle focused on a set of far-right Australian Facebook groups identified with a digital ethnographic study (for Facebook), the Twitter commercial API (for Twitter), and the Youtube API (for Youtube) using the following keywords as input: **bushfire, australian fires, arson, scottyfrommarketing, liarfromtheshiar, australiaburns, australiaburning, itsthegreensfault, backburning, back burning, climate change, climate emergency, climate hoax, climate crisis, climate action now**. It is important to point out that the Facebook

sample is sourced predominantly from far-right groups, whereas the Twitter and Youtube are general scrapes. Two sets of augmentations were added to the postings: the topic and the opinion of the post, obtained using a set of topic and opinion classifiers trained in [88]. The set of opinions was constructed via a qualitative study.

A limitation of the original SocialSense dataset is that the Twitter dataset for the Australian bushfires/ climate change topic was scraped only from December 2019 to February 2020, which did not capture early opinion during the start of the bushfire crisis. To that end, we decided to rescrape the Twitter dataset from November 1, 2019 to January 29, 2020 using the Twitter Academic v2 API and the same set of keywords. Since the Twitter Academic API does not allow querying based on user account location, we utilized AWS's Amazon Location Service to geocode users based on their free-text location and description fields and filtered only for tweets from Australian users. Finally, we applied the same set of topic and opinion classifiers to augment the Twitter data.

Once we aligned the Facebook dataset from SocialSense and the re-scraped Twitter dataset on the target timeframe, we observed that 10 (out of 34) opinions account for most of the Twitter (95%) and Facebook (81%) postings. To limit the set of opinions in our analysis, we focus on six **opinions of interest** constructed by merging subsets of the 10 opinions, after that, we filter the Twitter and Facebook datasets on this set of opinions. We index the six opinions we consider as {0, 1, 2, 3, 4, 5} and are shown below:

0. Greens influence and policy are the cause of the 2019-2020 Australian bushfires./ I am opposed to the policies of Greens political parties.
1. Mainstream media cannot be trusted.
2. Climate change crisis isn't real/ Climate change is a UN hoax/ Climate change is a scam to generate profit for the wealthy and powerful.
3. 2019-2020 Australian bushfires and climate change not related.
4. 2019-2020 Australian bushfires were caused by random arsonists.
5. Changes in the earth's climate are a natural, normal phenomenon/Bush fires are a normal summer occurrence for Australia.

Lastly, keeping in mind our goal of uncovering the interactions between sympathizers and opponents of the aforementioned problematic opinions, we furthermore differentiate whether the expressed opinion shows a far-right or moderate stance, which effectively splits our set of 6 opinions into 12 **stanced** opinions. For instance, the anti-Greens opinion (labeled 0) splits as far-right (labeled 0+) and moderate (labeled 0+). We represent our set of opinions as $\{(i-, i+) | i \in \{0, \ldots, 5\}\}$. We leverage the far-right stance detector introduced by Ram et al. [114] and apply it on each post of the aligned Facebook and Twitter dataset. In summary, the Bushfire Opinions dataset consists of posts on $P = 2$ platforms: 474,461 on Twitter and 27,974 on Facebook, exhibiting $M = 12$ stanced opinions. For compatibility with our discrete-time model, we aggregate post volumes on Facebook and Twitter into hourly counts, yielding $T = 2,160$ time points over the 90 days of November 1, 2019 to January 29, 2020.

4.5.3 VEVO 2017 Top 10 dataset

We assemble the VEVO 2017 Top 10 dataset by aligning artist-level time series of YouTube views and Twitter post counts ($P = 2$) for the top $M = 10$ VEVO-affiliated artists over $T = 100$ days from Jan 2, 2017 to Apr 11, 2017.

The YouTube time series are obtained from the **VEVO Music Graph dataset** [73], containing daily view counts for music videos posted by verified VEVO artists in six English-speaking countries (USA, UK, Canada, Australia, New Zealand, and Ireland). We combine the view counts for all music videos that belong to a given artist to obtain artist-level YouTube view time series. For Twitter, we leverage the Twitter API to get daily counts of posts with text containing an input query. We obtain the artist-level Twitter post time series using the artist's name as the input query.

Unlike the single exogenous signal $S(t)$ in the Bushfire Opinions dataset, we use a different exogenous signal $S_i(t)$ for each artist i – the Google Trends for each artist i. Using the set $\{S_i(t)\}$ instead of a single $S(t)$ requires several small changes to Eq. (4.5), Eq. (4.10), and the model gradients. We fully detail these changes in the online appendix [105]. We do not consider any interventions $\{X_k(t)\}$ as we seek to uncover endogenous interactions across artists.

4.6 Predictive Evaluation

This section evaluates the OMM's predictive capabilities on two real-world datasets. We introduce our prediction task, evaluation metrics and baselines, then present the results.

Model Setup. We use a temporal holdout strategy similar to prior literature [61, 35, 111]: we fit OMM on $\mathcal{T}_{obs}$ and evaluate performance on $\mathcal{T}_{pred}$. Backtesting is another viable alternate evaluation approach; however, it is significantly more computationally intensive, and we prefer the temporal holdout. For the bushfire case study, $\mathcal{T}_{obs} = \{1, \ldots, 1800\}$ where time is in hours (i.e., days 1–75 of our period of interest) and $\mathcal{T}_{pred} = \{1801, \ldots, 2160\}$ (i.e., days 76–90). For the VEVO case study, $\mathcal{T}_{obs} = \{1, \ldots, 75\}$ and $\mathcal{T}_{pred} = \{76, \ldots, 100\}$.

We consider two tasks: (1) opinion volume prediction and (2) opinion share prediction. For the first task, we predict the total volume of opinionated posts on the P platforms during the evaluation period. We measure performance using the platform-averaged symmetric mean absolute percentage error (SMAPE) of predicted volumes $\{\bar{n}_t^p | t \in \mathcal{T}_{pred}\}$ on platform p relative to the actual volumes $\{n_t^p | t \in \mathcal{T}_{pred}\}$,

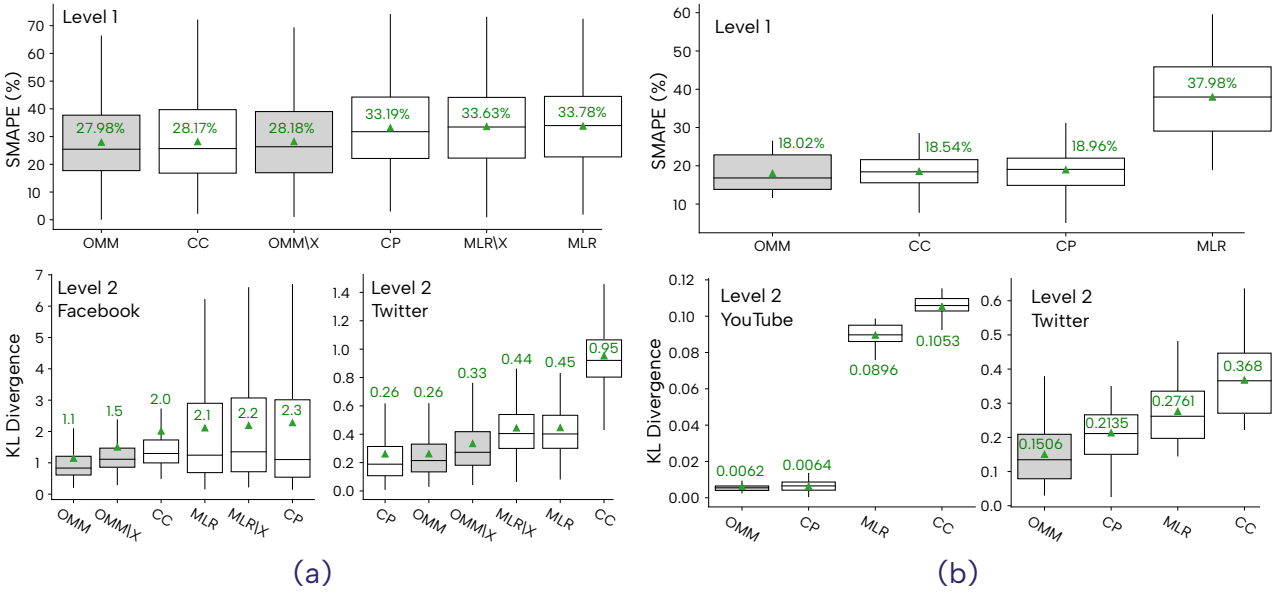
$$\text{SMAPE} = \frac{1}{P} \sum_{p=1}^{P} \left(\frac{100\%}{360} \sum_{t=1801}^{2160} \frac{|\bar{n}_t^p - n_t^p|}{|\bar{n}_t^p| + |n_t^p|} \right). \tag{4.11}$$

The predicted opinion volumes $\{\bar{n}_t^p\}$ are obtained using the OMM simulation algorithm. We (1) condition on $\{n_{i,t}^p | t \in \mathcal{T}_{obs}\}$, (2) run the algorithm to sample $\{n_{i,t}^p\}$ on $\mathcal{T}_{pred}$, then (3) sum over opinion types $\{i\}$ to get predicted opinion volumes np $n_t^p = \sum_i n_{i,t}^p$. We repeat $R = 5$ times, and average over the samples to obtain $\{\bar{n}_t^p | t \in \mathcal{T}_{pred}\}$.

For opinion share prediction, we predict the opinion market shares $\{s_{i,t}^p\}$ for each platform p on the evaluation period. To evaluate how well we predict opinion market shares, we calculate the KL divergence of predicted market shares $\{\bar{s}_t^p | t \in \mathcal{T}_{pred}\}$ (obtained similar to $\{\bar{n}_t^p\}$ described above) relative to actual market shares $\{s_t^p | t \in \mathcal{T}_{pred}\}$

$$\mathrm{KL}^p(t) = \sum_{i=1}^{M} s_{i,t}^p \log \frac{\bar{s}_{i,t}^p}{s_{i,t}^p}. \tag{4.12}$$

Figure 4.4: Predictive evaluation of OMM on (a) Bushfire Opinions and (b) VEVO 2017 Top 10 datasets. Boxplots are sorted left to right by the mean (shown with **green** triangle). Shaded boxplots correspond to versions of OMM. The top panels show the platform-averaged SMAPE of volumes on $\mathcal{T}_{pred}$. Bottom panels plot the KL divergence of predicted and actual market shares.



Baselines. We compare OMM with the discretized versions of the Correlated Cascades (CC) model [85] and Competing Products (CP) model [89] – the current state-of-the-art models in product share modeling, covered in related works. For the bushfire study, we test the effectiveness of interventions by fitting OMM without $\{X_k(t)\}$ (indicated as OMM\X).

We also consider a feature-based predictive baseline – the multivariate linear regression (MLR), used previously for online popularity prediction [115, 61]. We build MLR with a one-week sliding window of three types of features: the previous event counts, exogenous signal $S(t)$ and interventions $\{X_k(t)\}$. The predictive targets are the event counts $\{n^p_{i,t}\}$ for each point on $\mathcal{T}_{pred}$. Analogous to OMM fitted without interventions $\{X_k(t)\}$, we additionally train MLR without $\{X_k(t)\}$ (indicated as MLR\X) for the bushfire case study.

OMM, CC and CP are generative models typically designed for explainability and are known to be suboptimal for prediction [116]. In contrast, feature-driven approaches (e.g., MLR) use machine learning to predict using training features. Such approaches are designed mainly for prediction and have weaker explainability since they do not model the data-generation process [116]. In this work, we are interested in the dual tasks of predicting and explaining opinion market shares, hence our focus on generative approaches.

Predict Opinion Volumes. Fig. 4.3 (a) showcases the observed (blue line) and modelled (orange line) opinion volumes for the bushfire dataset. We visually observe that OMM achieves a tight fit on both the training and the prediction period (hashed area). The VEVO dataset results are shown in the online appendix [105]. We further compare OMM's predictive performances against baselines. The top row of boxplots in Fig. 4.4 (a) and Fig. 4.4 (b) shows the platform-averaged SMAPE of predicted volumes for the bushfire and VEVO datasets, respectively. We make two observations. First, in both case studies, OMM outperforms all baselines on opinion volume prediction. Second, OMM outperforms OMM\X, indicating the role of media coverage in shaping attention.

Predict Opinion Market Share. Fig. 4.3 (b) visualizes the observed (left column) and fitted during training and predicted during testing (right column) opinion market shares for the bush-fire dataset. We see that the opinion distribution on Twitter has significantly more variation than on Facebook and that OMM closely captures the trend in opinion shares on both platforms. The VEVO dataset results are in the online appendix [105]. Fig. 4.4 (a) and Fig. 4.4 (b) show the KL-divergence of predicted market shares for the bushfire (Facebook and Twitter) and VEVO (YouTube and Twitter) datasets, respectively. We make several observations. First, on the bushfire dataset, performance is better for Twitter than Facebook ($\mathrm{KL}^{TW}(t) < \mathrm{KL}^{FB}(t)$) due to Facebook having lower opinion counts than Twitter. Similarly, on the VEVO dataset $\mathrm{KL}^{YT}(t) < \mathrm{KL}^{TW}(t)$. Second, OMM consistently outperforms all baselines on both datasets, except for Twitter on bushfires, where CP and OMM are comparable. CC performs poorly since it does not model asymmetric opinion interactions and assumes all opinions reinforce or inhibit one another. CP performs poorly on Facebook (Twitter) for the bushfire (VEVO) dataset due to CP not having the notion of limited total attention. Due to higher bushfire postings on Twitter, CP pays more attention to Twitter. Lastly, OMM with $\{X_k(t)\}$ outperforms OMM without $\{X_k(t)\}$ on the bushfire dataset, suggesting that mainstream and controversial media effectively shape the opinion ecosystem.

4.7 Interpreting OMM Elasticities

In this section, we leverage the fitted OMM to uncover interactions across opinions and platforms in the bushfire dataset and artists in the VEVO dataset.

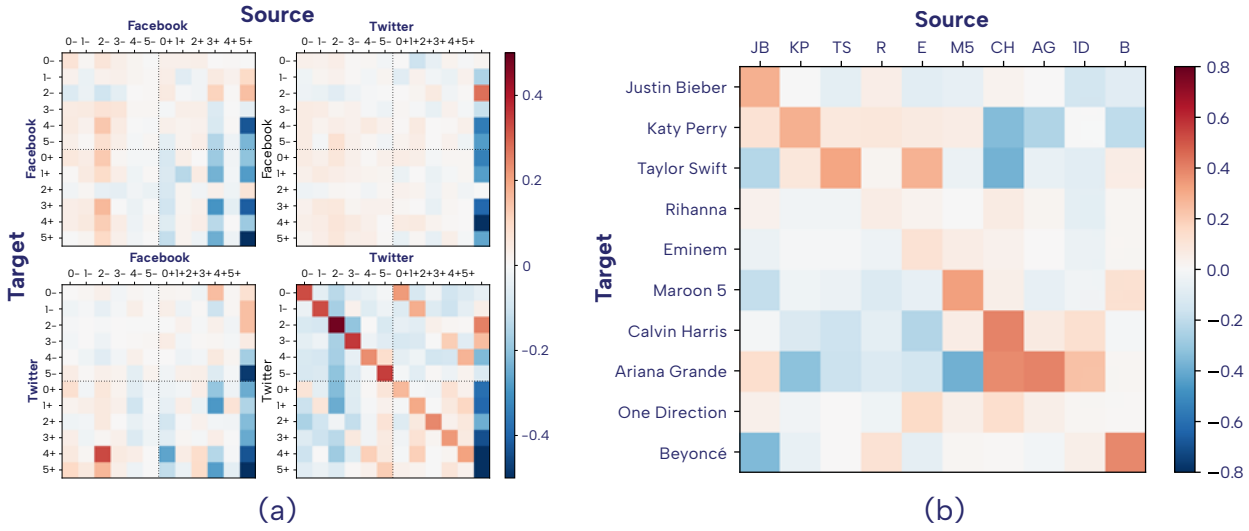
Uncovering Opinions Interactions. To study opinion interactions in the bushfire dataset, we calculate the opinion share model elasticities (see Eq. (4.4)) accounting for the endogenous volume $\lambda^p(t|j)$ and the intervention $\bar{X}_k(s)$ (see Eq. (4.9)). The endogenous elasticities $e(s_i^p(t), \lambda^q(t|j))$ quantify the competition-cooperation interactions across opinions. The intervention elasticity $e(s_i^p(t), \bar{X}_k(t))$ quantifies the sensitivity of opinion market shares to intervention $X_k(t)$. We derive the elasticities and show results for $e(s_i^p(t), \bar{X}_k(t))$ in the online appendix [105]. Fig. 4.5 (a) reports the time averages of $e(s_i^p(t), \lambda^q(t|j))$.

First, we study intra-platform reinforcement (top-left & bottom-right in Fig. 4.5 (a)). We see different behaviors for Facebook and Twitter. For Twitter, we have two observations. First, there is strong self-reinforcement for opinions (i.e., main diagonal), indicative of the echo chamber effect [90]. Second, there is significant cross-reinforcement among far-right sympathizers and opponents (i.e., diagonals on the upper-right & lower-left submatrices), implying exchanges or arguments between opposing camps. For Facebook, OMM detects little interaction among opinions, aside from the generally inhibitory effect of the opinions "Australian bushfires and climate change are unrelated" (3+) and "Bushfires are a normal summer occurrence" (5+) on other opinions. This is because Facebook data was collected from far-right groups with limited interaction with users of the opposing side.

How to Effectively Suppress Far-Right Opinions. The above implies that confrontation is not the most effective method to suppress far-right opinions, as it has the potential to backfire by bringing even more attention to them. A more effective method is boosting related counter-arguments; for instance, to suppress "Australian bushfires were caused by random arsonists" (4+) on Twitter, OMM indicates to promote "Climate change is real" (2-) and "Greens are not the cause of the bushfires" (0-). Boosting the opposite argument, i.e., "Australian bushfires were not caused by random arsonists" (4-), would backfire. The opinion "Bushfires are a normal summer occurrence in Australia" (5+) shows a different behaviour: it reinforces most moderate

opinions and inhibits far-right opinions. In particular, the "Bushfires are normal" opinion (5+) appears to trigger "Climate change is real" (2-), probably due to the diametric opposition nature of these opinions. The effect of 5+ on 2- holds across every pair of platforms. Additionally, on Facebook, "Australian bushfires and climate change are not related" (3+) has a similar effect on other opinions as the "Bushfires are normal" opinion (5+), probably due to the similarity of their topic content.

Figure 4.5: Interpretability of OMM. (a) Endogenous elasticities $e(s_i^p(t), \lambda^q(t|j))$ across opinion pairs (i, j) on respective platforms (p, q) in the bushfire dataset. Elasticities have direction and should be read from column (source) to row (target) for the platform and within each matrix. For example, the bottom-right matrix corresponds to influences from Twitter to Twitter; the cell {4-, 4+} ({row, column}) is the influence of opinion 4+ on 4-, positive and large meaning that 4+ has a strong reinforcing effect on 4-. (b) YouTube elasticities $e(s_i^{YT}(t), \lambda^{YT}(t|j))$ across artist pairs (i, j) in the VEVO 2017 Top 10 dataset.



Cross-Platform Reinforcement is generally weak due to the Facebook far-right groups acting as a filter bubble. Apart from the effect of "Bushfires are normal" (5+) (see above), there is little cross-reinforcement among opinions from Twitter to Facebook. In the bottom-left matrix of Fig. 4.5 (a), we see that "Australian bushfires and climate change are not related" (3+) affects other opinions in a similar way to "Bushfires are normal" (5+); furthermore, "Climate change is real" (2-) triggers "Australian bushfires were caused by arsonists" (4+).

Interactions Across VEVO Artists. Lastly, in Fig. 4.5 (b), we shift our attention to the VEVO dataset and look at the YouTube-to-YouTube elasticities $e(s_i^{YT}(t), \lambda^{YT}(t|j))$ across our set of artists. The Twitter and cross-platform elasticities are available in the online appendix [105].

We highlight three key observations. First, there is strong self-reinforcement for most artists (i.e., the main diagonal), an intuitive result reflecting these popular artists' strong fanbase. Second, OMM picks up non-trivial artist interactions that correspond with real-world events – the animosity and friendship relations show up in their popularity dynamics. For instance, we see that Calvin Harris inhibits both Taylor Swift (the two broke up in 2016[1]) and Katy Perry (the two had a long-lasting feud[2], due to Harris pulling out of Perry's 2011 tour last minute). Similarly, Taylor Swift and Justin Bieber have a mutually inhibiting relationship. The two have a well-

1 people.com/celebrity/taylor-swift-calvin-harris-breakup-timeline/

2 nme.com/news/music/katy-perry-ends-six-year-beef-calvin-harris-2128100

known uneasy relationship[1] since Justin Bieber and Selena Gomez used to date and the latter is one of Taylor Swift's close friends. Meanwhile, Calvin Harris and Ariana Grande have a reinforcing relationship, correlating with their collaboration "Heatstroke" released in March 2017. OMM picks up these relationships because we fit on online popularity driven by audience response. Fans of a given artist can choose to support or not support another artist based on real-world interactions, as indicated by the results above. Our third observation relates to the complexity of fanbase support for artists occupying the same genre: similar artists do not all just cooperate or compete for market share but can have unique pairwise relationships. For instance, Katy Perry, Taylor Swift and Ariana Grande occupy a similar niche (mainstream pop). However, our model uncovers that Taylor Swift and Katy Perry reinforce each other, while these two inhibit (and are inhibited by) Ariana Grande.

4.8 OMM as a Testbed for Interventions

The interventions $\{X_k(t)\}$ can lead to delayed effects in the opinion ecosystem due to the opinion dependency structure. For example, if an intervention is designed to boost a target opinion, it will indirectly boost all other opinions with a cooperative relationship with the target opinion. Furthermore, it will inhibit opinions with a competitive relationship with the target. Since elasticities only inform us of the instantaneous effect on opinion market shares, we perform a what-if exercise to study the role of interventions in the bushfire case study. We vary the intervention size and synthetically sample outcomes to observe the long-term effects of media coverage on the ecosystem.

What-if can inform A/B test design. We train OMM on observational data; therefore, the inferred effects of interventions $\{X_k(t)\}$ are not causal impact estimates but rather evidence of causal effects. However, the previous section demonstrates that OMM can uncover complex relationships across opinions, providing compelling evidence that OMM is also able to uncover relationships between opinions and interventions. Therefore, the what-if exercise in this section showcases OMM as a testbed for interventions, usable for designing A/B testing that determines true causal effects. The OMM informs us of the effectiveness of interventions, allowing us to prioritize which specific interventions to test.

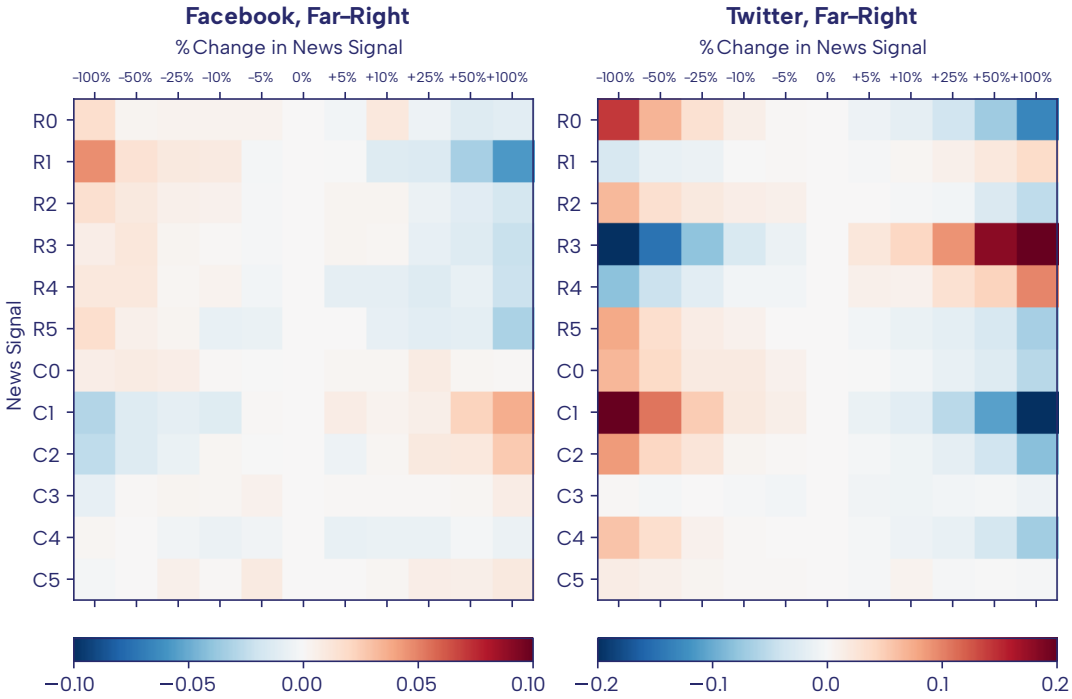
"What-if" Setup. We test the effect of interventions by synthetically increasing or decreasing their volumes past a given time point (see top panel of Fig. 4.1) and measuring the percentage change in far-right opinions. Let $k^* \in \{1, \ldots, K\}$ be the index of the modulated intervention. We modulate $X_{k^*}(t)$ as $X^{(r)}_{k^*}(t) = X_{k^*}(t) + r \cdot \mu_{X_{k^*}} \cdot 1_{(t>1800)}$, where $1_{(\cdot)}$ is the indicator function and $\mu_{X_{k^*}}$ is the mean volume of $X_{k^*}(t)$ on $\mathcal{T}_{obs}$. The parameter r controls the percent increase ($r > 0$) or decrease ($r < 0$) in media coverage beyond the change point $t = 1800$; $r = 0$ is the original $X_{k^*}(t)$. We run OMM with $X^{(r)}_{k^*}(t)$ for various r, and keep $X_k(t)$ fixed for $k \neq k^*$. We quantify the effects of intervention $X_{k^*}(t)$ as the average percent change (relative to $r = 0$) in the opinion market shares after the change point, i.e., $\mathcal{T}_{pred}$. We perform this procedure for all $k^* \in \{1, \ldots, K\}$.

How News Influences Far-Right Opinions. Fig. 4.6 shows the average percent changes in the market share of far-right opinions when modulating the interventions $\{R_i(t), C_i(t)\}$ one at a time for various r over 50 simulations. On Facebook, far-right opinions are suppressed by reputable news and reinforced by the majority of controversial news, except for news concerning "Greens policies are the cause of the Australian bushfires" (R_0) and "Australian bushfires were caused by arsonists" (R_4). On Twitter, both reputable and controversial news suppress far-right opinions, except for reputable news concerning "Australian bushfires/climate change

1 people.com/music/justin-bieber-selena-gomez-relationship-look-back/

are unrelated" (R_3), "Australian bushfires were caused by arsonists" (R_4) and to a lesser extent "Mainstream media cannot be trusted" (R_1).

Figure 4.6: We modulate the volume of reputable (R) and controversial (C) news for each opinion (in {0, 1, 2, 3, 4, 5}) from −100% to 100% of the mean volume and simulate OMM to see the percent change in the far-right (+) opinion market shares on Facebook (left) and Twitter (right).



We have two key insights. First, we see that the effect of the news on Facebook is modest compared to Twitter since the far-right public groups on Facebook behave as almost perfect filter bubbles in which news has little penetration. Second, indiscriminately increasing reputable news is not an effective strategy for suppressing far-right opinions on Twitter (see R_3 and R_4). Doing so can backfire since it brings even more attention to far-right users and their narratives [117].

How to effectively use the testbed. Assuming that A/B testing is performed by an entity in control of reputable news coverage (R_i here above), the results above indicate that the test should mainly concentrate on the effects of increasing R_1 (on Facebook), increasing R_0 and decreasing R_3 and R_4 (on Twitter). We leave as future work the design and execution of such an experiment. Our analysis in this chapter focuses on mitigating far-right opinions with media coverage. However, OMM can be leveraged as an intervention evaluation tool for information operations in other domains and fighting mis- & disinformation and online propaganda.

4.9 Summary and Discussion

This work introduces the Opinion Market Model (OMM), a novel two-tier model of the dynamics of the online opinion ecosystem. The first tier models the size of the attention market, and the second tier models opinions competing or cooperating for limited public attention under the influence of positive interventions. We develop algorithms to simulate and estimate OMM, showing the convergence using synthetic data. We demonstrate real-world applicability on a dataset of Facebook and Twitter discussions containing moderate and far-right opinions on bushfires and climate change [88] and a dataset of YouTube and Twitter attention

volumes for popular artists on VEVO [73]. We show OMM predicts opinion market shares better than state-of-the-art baselines [89, 85] and uncovers latent competitive and cooperative interactions across opinions: self-reinforcement attributable to the echo chamber effect and interactions between far-right sympathizers and opponents. Lastly, we quantify the effect of reputable and controversial media coverage on Facebook and Twitter. Scope of Study. This work focuses on the manifestation of far-right opinions in the context of the 2019-2020 Australian bushfires. Note that far-right ideology manifests in other political issues (e.g., gun control, LGBT rights, xenophobia), which we do not tackle here. Moreover, we do not focus on the general political science of far-right ideology since we are projecting onto a specific context.

5. Prototype Software to Monitor Online Social Media Discussions and Ads Collection

5.1 The Misinformation Dashboard

The following section provides an overview of the information dashboard developed by the Behavioral Data Science lab at the University of Technology, Sydney. The software is a comprehensive tool for collecting and analyzing data from social media and traditional media sources, focusing on specific topics of interest. This section summarizes the key features and functionalities of the dashboard and outlines our plans for further development.

We have conducted a live demonstration and recorded a video showcasing the functionality of the Misinformation Dashboard. The video can be accessed at the following link: https://youtu. be/DcW qNJ1UFg. Additionally, the dashboard can be accessed directly at https://dashboard.behavioral-ds.science.

5.1.1 Dashboard Features

The Misinformation Dashboard is designed to facilitate data collection from both social media and traditional media platforms regarding specific topics of interest. Users can configure multiple topics of interest on the left side of the interface. To configure a new topic, users must define its name and specify relevant keywords. The dashboard allows specifying excluded keywords to offer greater flexibility in refining search queries.

Data Collection Process. Once a search query is defined, the dashboard connects to the APIs of social media platforms such as Instagram, Facebook, and Twitter, as well as a wide selection of traditional media sources. It then initiates the data collection process to gather data from these sources.

Labeling System. The dashboard includes a labeling system to categorize the collected data. The labeling feature enables the definition of two types of labels: opinions and themes. Opinions represent viewpoints expressed by individuals who can support the opinion (agreement) or disagree. Themes refer to the main topics of discussion within social media. The labeling system allows users to code opinions and themes through a codebook. The video demonstrates the process of adding new labels and the ability to describe each label and build a comprehensive codebook.

Data Analysis and Labeling. The Information Dashboard provides various options for analyzing the collected data. Users can access detailed information about individual posts, including the full text, source platform, URL, and metadata. Moreover, users can start adding labels to subsets of social media data, indicating agreement or disagreement with specific opinions and association with particular themes. This allows for targeted labeling and segmentation of the dataset.

Machine Learning Integration. While currently under development, the dashboard aims to incorporate machine learning capabilities. The goal is to train classifiers in the background using the labeled data to automatically label all posts within the dataset. By leveraging advanced AI tools, such as generative Large Language Models (LLMs) like ChatGPT, the dashboard aims to automate the labeling process. Notably, the human expert can correct machine-assigned labels, enabling the system to learn from these corrections and improve over time. The codebook's vocabulary can also evolve by adding new opinions as they emerge during the dashboard's operation.

5.1.2 Further Development Plans

We outline the following plans for the further development and enhancement of the Misinformation Dashboard:

1. Implement human-in-the-loop auto-labeling of posts: This feature will utilize Natural Language Processing (NLP) techniques to train classifiers that can automatically label posts based on the predefined topics and themes within the dashboard. The goal is to involve human experts in labeling to refine the algorithm's accuracy. We have conducted peer-reviewed research on this technology [88], demonstrating performance accuracy of up to 80%.
2. Implement extremist user ideology detection: A novel tab will be introduced to study users engaged in monitored discussions. By utilizing the concepts of user homophily and psycholinguistic patterns [114], we aim to create an early detection system capable of identifying language patterns that may be indicative of extreme-leaning discourse

5.1.3 Merits

The Information Dashboard prototype offers several merits, including:

- Professional software development: The prototype is developed by RAPIDO, UTS's professional software development team. RAPIDO has extensive software development expertise for government, defense, and industrial partners.
- State-of-the-art research integration: The prototype will incorporate cutting-edge research results, ensuring the utilization of techniques and methodologies that have demonstrated high performance in capturing and analyzing misinformation-related data.

5.1.4 Conclusion

The Information Dashboard provides a comprehensive tool for capturing and analyzing data from both social media and traditional media sources. The outlined plans for further development, which include human-in-the-loop auto-labeling and extremist user ideology detection, demonstrate the commitment to leveraging advanced technologies and research findings to enhance the prototype's effectiveness. The collaboration between UTS and the stakeholders ensures a user-centered approach that aligns with the specific requirements of the stakeholders.

5.2 Advertisements: who is funding misleading articles?

Online misinformation is fuelled financially by advertisement revenue. Advertisements are delivered in an online ecosystem through ad networks and ad exchanges, which take on the task of connecting advertisers (independent parties) to suppliers (websites). These ad networks financially incentivise website hosts to place ads on their websites. The more traffic the host can attract to their website, the more money they can make. Many misinformation websites are created explicitly to farm ad revenue, with societal harm written off as a necessary by-product.

Despite great pressure on the largest ad networks and exchanges, little action has been taken in pursuit of moderation. Any website host, regardless of their content, can implement advertisements without oversight from the ad networks themselves. Often this results in ad content from legitimate companies being placed near misinformation content. The companies leveraging these ad networks for delivery are marketing their brand, and this goal is not supported by associating the brand with harmful content.

We take a new approach to attacking misinformation financially. Given the inaction by ad networks to do even the most essential content moderation, we go downstream and look to the companies advertising these networks to put pressure on misinformation website hosts and the ad networks that support them. To motivate this pressure, we must demonstrate to the advertisers that their brands are associated with harmful content. To gather this data, we built an advertisement scraper.

5.2.1 Data collection

Given a set of misinformation websites, we are interested in gathering the advertisements displayed on them regularly and certain descriptive metadata about the conditions surrounding the ad delivery. In this context, advertisements refer to the code and data that comprise an online ad. From this code and data, we can infer the company the advertisement is being delivered for.

When ad networks deliver ads, they do so in a targeted manner. Using the information they gather from your online activity as your ISP, they programmatically decide which ad will be delivered to your web browser. For example, a user in Melbourne is more likely to receive ads about local Melbourne businesses than businesses in Sydney. For each site we visit, we gather relevant information leveraged by the ad network to choose which advertisement we see.

In addition to this metadata, we collect the URL the advertisement links to, often leading to the domain of the particular company providing the advertisement. This allows us to identify the origin of the advertisements. In some cases, the link is an alias and does not contain the domain name of the advertising company. For these situations, we run a secondary scrape on the link we gathered to collect metadata descriptive of the company.

We collect a screenshot from each website to verify that the ads we scraped match the ads on the page. These screenshots are a powerful visual we can provide advertisers to show concrete evidence that their company was advertised alongside harmful content.

5.2.2 Technical details

The web scraper is implemented using OpenWPM, a framework developed specifically for web privacy research. Using OpenWPM, we visit each site we're interested in and scrape the HTML recursively through iframes. An iframe (inline frame) is an HTML block containing another HTML document. Advertisements are frequently loaded as iframes which is why the recursive scrape is important. After each scrape, we have a full body of HTML from the top-level site as well as all the iframes.

From within this HTML, we identify and extract information relevant to the underlying content of the advertisements, ultimately producing a list of advertised URLs associated to each top-level site. This data is aggregated in an SQLite database. To ensure we get the maximum information about each ad, we take the URLs we scraped from the advertisement and set off another scraper on each of those sites. From this scrape, we gather metadata which is aggregated in another table. See Tables 5.1 and 5.2 for a schema of this database.

Table 5.1: Data collected from top level scrape.

field	type	description
id	integer	ID of scrape.
date	text	Date and time of scrape.
name	text	Name of site being scraped.
source file	text	Location of scraped HTML.
screenshot file	text	Location of screenshot of site.
backend mobile detect	text	utag. Whether the site has been visited on mobile.
backend geo country	text	utag. Country from which site was visited.
backend geo region	text	utag. Region from which site was visited.
backend geo city	text	utag. City from which site was visited.
backend geo lat	float	utag. Latitude from which site was visited.
backend geo long	float	utag. Longitude from which site was visited.
backend geo tmz	text	utag. Time zone from which site was visited
backend geo network	text	ISP network from which site was visited
mfSponsor	text	Advertisement URL
p tags	text	Text about underlying advertisement.
adurls	text	Advertisement URL
destinationUrl	text	Advertisement URL

Table 5.2: Data collected from advertisements found on top-level sites.

field	type	description
id	int	ID of secondary scrape
scrape id	int	Foreign key, references “id” in other table
name	text	Name of site from top level scrape
date	text	Date and time of secondary scrape
url	text	URL of site being scraped
filename	text	Location where HTML is stored
keywords	text	Content in “keywords” meta tag
description	text	Content in “description” meta tag
title	text	Content in “title” meta tag
og title	text	Content in “og title” meta tag
og site name	text	Content in “og site name” meta tag
og description	text	Content in “og description” meta tag
twitter keywords	text	Content in “twitter keywords” meta tag
twitter description	text	Content in “twitter description” meta tag
twitter title	text	Content in “twitter title” meta tag
twitter site	text	Content in “twitter site” meta tag

A full technical overview of the web scraper is available on GitHub.

Bibliography

[1] Cindy K Chung and James W Pennebaker. What do we know when we liwc a person? text analysis as an assessment tool for traits, personal concerns and life stories. The Sage handbook of personality and individual differences, pages 341–360, 2018.

[2] Lazar Stankov. From social conservatism and authoritarian populism to militant right-wing extremism. Personality and individual differences, 175:110733, 2021.

[3] Inez Okulska and Anna Zawadzka. Styles with benefits. the stylometrix vectors for stylistic and semantic text classification of small-scale datasets and different sample length. 2023.

[4] Isabelle van der Vegt, Maximilian Mozes, Bennett Kleinberg, and Paul Gill. The grievance dictionary: Understanding threatening language use. Behavior research methods, pages 1– 15, 2021.

[5] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. Journal of language and social psychology, 29(1):24– 54, 2010.

[6] Marcella Andrade da Rocha, Philippi Sedir Grilo de Morais, Daniele Montenegro da Silva Barros, João Paulo Queiroz dos Santos, Sara Dias-Trindade, and Ricardo Alexsandro de Medeiros Valentim. A text as unique as a fingerprint: Text analysis and authorship recognition in a virtual learning environment of the unified health system in brazil. Expert Systems with Applications, 203:117280, 2022.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[8] Arie W Kruglanski, Erica Molinario, Molly Ellenberg, and Gabriele Di Cicco. Terrorism and conspiracy theories: A view from the 3n model of radicalization. Current Opinion in Psychology, pages 101–396, 2022.

[9] Jytte Klausen, Selene Campion, Nathan Needle, Giang Nguyen, and Rosanne Libretti. Toward a behavioral model of “homegrown” radicalization trajectories. Studies in Conflict & Terrorism, 39(1):67–83, 2016.

[10] Ahmad El-Muhammady. Radicalisation model: Learning from malaysian militant–extremists. In Terrorist Deradicalisation in Global Contexts, pages 155–184. Routledge, 2019. 76

[11] David A Winter and Guillem Feixas. Toward a constructivist model of radicalization and deradicalization: A conceptual and methodological proposal. Frontiers in psychology, 10(412):1–11, 2019.

[12] Clark McCauley and Sophia Moskalenko. Mechanisms of political radicalization: Pathways toward terrorism. Terrorism and political violence, 20(3):415–433, 2008.

[13] Bertjan Doosje, Fathali M Moghaddam, Arie W Kruglanski, Arjan De Wolf, Liesbeth Mann, and Allard R Feddes. Terrorism, radicalization and de-radicalization. Current Opinion in Psychology, 11:79–84, 2016.

[14] Bergeron Catherine and Brunet Louis. A path of radicalization: Complementary of the group and the individual flaw. Journal of Forensic Psychology Research and Practice, pages 178–200, 2022.

[15] Rosemary Pepys, Robert Bowles, and Noémie Bouhana. A simulation model of the radicalisation process based on the ivee theoretical framework. Journal of Artificial Societies and Social Simulation, 23(3):1–22, 2020.

[16] Matteo Vergani, Muhammad Iqbal, Ekin Ilbahar, and Greg Barton. The three ps of radicalization: Push, pull and personal. a systematic scoping review of the scientific evidence about radicalization into violent extremism. Studies in Conflict & Terrorism, 43(10):854–854, 2020.

[17] Jan-Jaap Van Eerten and Bertjan Doosje. Challenging extremist views on social media: Developing a counter-messaging response. Routledge, 2019.

[18] Joshua Thorburn. The (de-) radical (-ising) potential of r/incelexit and r/exredpill. European Journal of Cultural Studies, 26(3):464–471, 2023.

[19] Rachel Martin. Former Anti-Vaccine Mom Explains How Movement Pulled Her In, And How She Left npr, 2021.

[20] Bo Hamby Steve Mullis, Rachel Martin. She Resisted Getting Her Kids The Usual Vaccines. Then The Pandemic Hit rolling stone, 2021.

[21] Melody Schreiber. 'The only logical choice': anti-vaxxers who changed their minds on Covid vaccines the guardian, 2022.

[22] April Rose. 'Brainwashed' mum refused to vax kids news.com, 2021.

[23] Anastasiia Carrier. QAnon Almost Destroyed My Relationship. Then My Relationship Saved Me From QAnon. politico, 2021.

[24] Jake Rosen. Former QAnon believer says following the conspiracy "was absolutely a drug" cbs news, 2021.

[25] EJ Dickson. How Ex-QAnon Followers Escaped The Cultish Conspiracy Theory rolling stone, 2021.

[26] Molly Boigon. She was a Jewish QAnon supporter. And she warns it could happen to you - U.S. News haaretz, 2021.

[27] Donie O'Sullivan. She was stunned by Biden's inauguration. How this South Carolina mom escaped QAnon cnn, 2021.

[28] Chris Stewart. Former anti-vaccine influencer now using social media to encourage others to get vaccinated denver 7, 2021.

[29] Craig Idlebrook. I was once a hardcore anti-vaxxer. Now I try to nudge people to get the Covid-19 vaccine start news, 2021.

[30] Jean Burgess and Nancy K Baym. Twitter: A Biography. NYU Press, New York, NY, 2020.

[31] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. Science, 359(6380):1146–1151, 2018.

[32] Marian-Andrei Rizoiu, Timothy Graham, Rui Zhang, Yifei Zhang, Robert Ackland, and Lexing Xie. #DebateNight: The role and influence of socialbots on Twitter during the 1st 2016 U.S. presidential debate. In Proceedings of the Twelfth International AAAI Conference on Web and Social Media, pages 300–309, 2018.

[33] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In Proceedings of the International AAAI Conference on Web and Social Media, volume 11, pages 512–515, 2017.

[34] Digital Services Act. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act). OJ, L 277:1–102, 2022.

[35] Marian-Andrei Rizoiu, Swapnil Mishra, Quyu Kong, Mark Carman, and Lexing Xie. SIR-Hawkes: Linking epidemic models and Hawkes processes to model diffusions in finite populations. In Proceedings of the 2018 World Wide Web Conference, pages 419–428, 2018.

[36] Scott M. Graffius. Lifespan (half-life) of social media posts: Update for 2023, 2023. (accessed 28 March 2023).

[37] Albert-László Barabási. The origin of bursts and heavy tails in human dynamics. Nature, 435(7039):207–211, 2005.

[38] Kristina Lerman and Rumi Ghosh. Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. In Proceedings of the International AAAI Conference on Web and Social Media, volume 4, pages 90–97, 2010.

[39] Fei Xiong and Yun Liu. Opinion formation on social media: An empirical approach. Chaos: An Interdisciplinary Journal of Nonlinear Science, 24(1):013130, 2014.

[40] David R Bild, Yue Liu, Robert P Dick, Z Morley Mao, and Dan S Wallach. Aggregate characterization of user behavior in Twitter and analysis of the retweet graph. ACM Transactions on Internet Technology (TOIT), 15(1):1–24, 2015. 78

[41] Peter Mathews, Lewis Mitchell, Giang Nguyen, and Nigel Bean. The nature and origin of heavy tails in retweet activity. In Proceedings of the 26th International Conference on World Wide Web Companion, pages 1493–1498, 2017.

[42] Riley Crane and Didier Sornette. Robust dynamic classes revealed by measuring the response function of a social system. Proceedings of the National Academy of Sciences, 105(41):15649–15653, 2008.

[43] Yasuko Matsubara, Yasushi Sakurai, B Aditya Prakash, Lei Li, and Christos Faloutsos. Rise and fall patterns of information diffusion: model and implications. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 6–14, 2012.

[44] Marian-Andrei Rizoiu, Young Lee, Swapnil Mishra, and Lexing Xie. Hawkes processes for events in social media. In Frontiers of Multimedia Research, pages 191–218. ACM, New York, NY, 2017.

[45] Rafael Lima. Hawkes processes modeling, inference, and control: An overview. SIAM Review, 65(2):331–374, 2023.

[46] Alan G Hawkes. Hawkes processes and their applications to finance: A review. Quantitative Finance, 18(2):193–198, 2018.

[47] Alex Reinhart. A review of self-exciting spatio-temporal point processes and their applications. Statistical Science, 33(3):299–318, 2018.

[48] Oleksandr Shchur, Ali Caner Türkmen, Tim Januschowski, and Stephan Günnemann. Neural temporal point processes: A review. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, pages 4585–4593, 2021.

[49] Oliver Milman. #ClimateScam: Denialism claims flooding Twitter have scientists worried. Guardian, 2022. (accessed 28 March 2023).

[50] Darren L Linvill and Patrick L Warren. Troll factories: Manufacturing specialized disinformation on Twitter. Political Communication, 37(4):447–467, 2020.

[51] Alan G Hawkes. Spectra of some self-exciting and mutually exciting point processes. Biometrika, 58(1):83–90, 1971.

[52] Swapnil Mishra, Marian-Andrei Rizoiu, and Lexing Xie. Feature driven and point process approaches for popularity prediction. In Proceedings of the 25th ACM International Conference on Information and Knowledge Management, pages 1069–1078, 2016.

[53] Michael Mark and Thomas A Weber. Robust identification of controlled Hawkes processes. Physical Review E, 101(4):043305, 2020.

[54] Yosihiko Ogata and Jiancang Zhuang. Space–time ETAS models and an improved extension. Tectonophysics, 413(1-2):13–23, 2006.

[55] Minkyoung Kim, Dean Paini, and Raja Jurdak. Modeling stochastic processes in disease spread across a heterogeneous social system. Proceedings of the National Academy of Sciences, 116(2):401–406, 2019.

[56] Andrea L Bertozzi, Elisa Franco, George Mohler, Martin B Short, and Daniel Sledge. The challenges of modeling and forecasting the spread of COVID-19. Proceedings of the National Academy of Sciences, 117(29):16732–16738, 2020.

[57] Mehrdad Farajtabar, Yichen Wang, Manuel Gomez-Rodriguez, Shuang Li, Hongyuan Zha, and Le Song. COEVOLVE: A joint point process model for information diffusion and network evolution. Journal of Machine Learning Research, 18(41):1–49, 2017.

[58] George O Mohler, Martin B Short, P Jeffrey Brantingham, Frederic Paik Schoenberg, and George E Tita. Self-exciting point process modeling of crime. Journal of the American Statistical Association, 106(493):100–108, 2011.

[59] Peng Bao, Hua-Wei Shen, Xiaolong Jin, and Xue-Qi Cheng. Modeling and predicting popularity dynamics of microblogs using self-excited Hawkes processes. In Proceedings of the 24th International Conference on World Wide Web, pages 9–10, 2015.

[60] Ryota Kobayashi and Renaud Lambiotte. TiDeH: Time-dependent Hawkes process for predicting retweet dynamics. In Proceedings of the International AAAI Conference on Web and Social Media, volume 10, pages 191–200, 2016.

[61] Marian-Andrei Rizoiu, Lexing Xie, Scott Sanner, Manuel Cebrian, Honglin Yu, and Pascal Van Hentenryck. Expecting to be HIP: Hawkes intensity processes for social media popularity. In Proceedings of the 26th International Conference on World Wide Web, pages 735–744, 2017.

[62] Emmanuel Bacry, Iacopo Mastromatteo, and Jean-François Muzy. Hawkes processes in finance. Market Microstructure and Liquidity, 1(01):1550005, 2015.

[63] Massil Achab, Emmanuel Bacry, Jean-François Muzy, and Marcello Rambaldi. Analysis of order book flows using a non-parametric estimation of the branching ratio matrix. Quantitative Finance, 18(2):199–212, 2018.

[64] Michael Mark, Jan Sila, and Thomas A Weber. Quantifying endogeneity of cryptocurrency markets. European Journal of Finance, 28(7):784–799, 2022.

[65] Andrew Daw and Jamol Pender. Queues driven by Hawkes processes. Stochastic Systems, 8(3):192–229, 2018.

[66] Alexander Wehrli and Didier Sornette. The excess volatility puzzle explained by financial noise amplification from endogenous feedbacks. Scientific Reports, 12(1):18895, 2022.

[67] Thierry Bochud and Damien Challet. Optimal approximations of power laws with exponentials: Application to volatility models with long memory. Quantitative Finance, 7(6):585– 589, 2007.

[68] Stephen J Hardiman, Nicolas Bercot, and Jean-Philippe Bouchaud. Critical reflexivity in financial markets: A Hawkes process analysis. European Physical Journal B, 86:1–9, 2013.

[69] Marcello Rambaldi, Paris Pennesi, and Fabrizio Lillo. Modeling foreign exchange market activity around macroeconomic news: Hawkes-process approach. Physical Review E, 91(1):012819, 2015.

[70] Paul M Barrett. Who moderates the social media giants? NYU Stern Center for Business & Human Rights, 2020.

[71] Samrat Gupta, Gaurav Jain, and Amit Anand Tiwari. Polarised social media discourse during covid-19 pandemic: evidence from youtube. Behaviour & Information Technology, pages 1–22, 2022.

[72] Utkarsh Upadhyay, Abir De, Aasish Pappu, and Manuel Gomez-Rodriguez. On the complexity of opinions and online discussions. WSDM '19, 2019.

[73] Siqi Wu, Marian-Andrei Rizoiu, and Lexing Xie. Estimating Attention Flow in Online Video Networks. CSCW, 3:1–25, nov 2019.

[74] Elmie Nekmat. Nudge effect of fact-check alerts: source influence and media skepticism on sharing of news misinformation in social media. Social Media+Society, 6(1), 2020.

[75] Greyson K Young. How much is too much: the difficulties of social media content moderation. Information & Communications Technology Law, 31(1):1–16, 2022.

[76] Sam Jackson. The double-edged sword of banning extremists from social media. 2019.

[77] Katherine Clayton, Spencer Blair, Jonathan A Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, et al. Real solutions for fake news? measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. Political Behavior, 42(4):1073–1095, 2020.

[78] Ethan Porter and Thomas J Wood. Fact checks actually work, even on facebook. but not enough people see them. The Washington Post, 2021.

[79] GIFCT. Content-sharing algorithms, processes, and positive interventions working group. 2021.

[80] Gulizar Haciyakupoglu, Jennifer Yang Hui, VS Suguna, Dymples Leong, and Muhammad Faizal Bin Abdul Rahman. Countering fake news: A survey of recent global initiatives. 2018.

[81] Kai Shu, Suhang Wang, and Huan Liu. Beyond news contents: The role of social context for fake news detection. WSDM '19, page 312–320, 2019.

[82] Minna Horowitz, Stephen Cushion, Marius Dragomir, Sergio Gutiérrez Manjón, and Mervi Pantti. A framework for assessing the role of public service media organizations in countering disinformation. Digital Journalism, 10(5), 2022. 81

[83] Courtney Radsch. Media development and countering violent extremism: An uneasy relationship, a need for dialogue. Center for International Media Assistance. (2016), 2016.

[84] Lee G. Cooper. Chapter 6 market-share models. In Marketing, volume 5, pages 259–314. Elsevier, 1993.

[85] Ali Zarezade, Ali Khodadadi, Mehrdad Farajtabar, Hamid R. Rabiee, and Hongyuan Zha. Correlated cascades: Compete or cooperate. AAAI 2017, pages 238–244, 2017.

[86] Lilian Weng, Alessandro Flammini, Alessandro Vespignani, and Fillipo Menczer. Competition among memes in a world with limited attention. Scientific reports, 2(1):335, 2012.

[87] Sarah Gelper, Ralf van der Lans, and Gerrit van Bruggen. Competition for attention in online social networks: Implications for seeding strategies. Management Science, 2021.

[88] Quyu Kong, Emily Booth, Francesco Bailo, Amelia Johns, and Marian-Andrei Rizoiu. Slipping to the extreme: A mixed method to explain how extreme opinions infiltrate online discussions. In Proceedings of the Sixteenth International AAAI Conference on Web and Social Media, pages 524–535. PMLR, May 2022.

[89] Isabel Valera and Manuel Gomez-Rodriguez. Modeling Adoption and Usage of Competing Products. In ICDM, November 2015.

[90] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. The echo chamber effect on social media. PNAS, 118(9), 2021.

[91] Abir De, Isabel Valera, Niloy Ganguly, Sourangshu Bhattacharya, and Manuel Gomez-Rodriguez. Learning and forecasting opinion dynamics in social networks. NIPS'16, pages 397–405, 2016.

[92] Bhushan Kulkarni, Sumit Agarwal, Abir De, Sourangshu Bhattacharya, and Niloy Ganguly. Slant+: A nonlinear model for opinion dynamics in social networks. In ICDM, pages 931– 936. IEEE, 2017.

[93] Tommaso Venturini and Richard Rogers. "api-based research" or how can digital sociology and journalism studies learn from the facebook and cambridge analytica data breach. Digital Journalism, 7(4):532–540, 2019.

[94] Michelle Betz. Constraints and opportunities: what role for media development in countering violent extremism? 2016.

[95] Kate Ferguson. Countering violent extremism through media and communication strategies. Reflections, 27:28, 2016.

[96] Adam Henschke and Alastair Reed. Toward an Ethical Framework for Countering Extremist Propaganda Online. Studies in Conflict & Terrorism, pages 1–18, March 2021.

[97] Marian-Andrei Rizoiu and Lexing Xie Xie. Online popularity under promotion: Viral potential, forecasting, and the economics of time. ICWSM, 11(1):182–191, May 2017.

[98] Kazuki Fujita, Alexey Medvedev, Shinsuke Koyama, Renaud Lambiotte, and Shigeru Shinomoto. Identifying exogenous and endogenous activity in social media. Phys. Rev. E, 98:052304, Nov 2018.

[99] Michele Garetto, Emilio Leonardi, and Giovanni Luca Torrisi. A time-modulated Hawkes process to model the spread of COVID-19 and the impact of countermeasures. 2021.

[100] Gary King, Benjamin Schneer, and Ariel White. How the news media activate public expression and influence national agendas. Science, 358:776–780, 2017.

[101] Andrew M. Guess, Pablo Barberá, Simon Munzert, and JungHwan Yang. The consequences of online partisan media. PNAS, 118(14), 2021.

[102] Massimiliano Agovino, Maria Rosaria Carillo, and Nicola Spagnolo. Effect of Media News on Radicalization of Attitudes to Immigration. Journal of Economics, Race, and Policy, December 2021.

[103] Raiha Browning, Deborah Sulem, Kerrie Mengersen, Vincent Rivoirard, and Judith Rousseau. Simple discrete-time self-exciting models can describe complex dynamic processes: A case study of COVID-19. PLoS ONE, 16, 2021.

[104] Alan G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. Biometrika, 58(1), 1971.

[105] Appendix: Opinion market model: Stemming far-right opinion spread using positive interventions, 2022. https://bit.ly/3LEokFT.

[106] Gianluca Detommaso, Hanne Hoitzing, Tiangang Cui, and Ardavan Alamir. Stein variational online changepoint detection with applications to hawkes processes and neural networks, 2019.

[107] Marian-Andrei Rizoiu, Alexander Soen, Shidi Li, Pio Calderon, Leanne Dong, Aditya Krishna Menon, and Lexing Xie. Interval-censored Hawkes processes. Journal of Machine Learning Research, 23(338), 2022.

[108] Rohit Ram, Emma Thomas, David Kernot, and Marian-Andrei Rizoiu. Detecting Extreme Ideologies in Shifting Landscapes: an Automatic & Context-Agnostic Approach, aug 2022.

[109] Karthik Sheshadri and Munindar P. Singh. The public and legislative impact of hyperconcentrated topic news. Science Advances, 5(8), 2019.

[110] Gabriel J Milinovich, Gail M Williams, Archie C A Clements, and Wenbiao Hu. Internet-based surveillance systems for monitoring emerging infectious diseases. The Lancet Infectious Diseases, 14(2):160–168, 2014.

[111] Quyu Kong, Marian-Andrei Rizoiu, and Lexing Xie. Describing and predicting online items with reshare cascades via dual mixture self-exciting processes. In 29th ACM CIKM, pages 645–654, 2020.

[112] Rajiv Johal. Factiva: Gateway to business information. Journal of Business & Finance Librarianship, 15(1):60–64, 2009.

[113] Maurício Gruppi, Benjamin D. Horne, and Sibel Adalı. Nela-gt-2019: A large multi-labelled news dataset for the study of misinformation in news articles, 2020.

[114] Rohit Ram, Emma Thomas, David Kernot, and Marian-Andrei Rizoiu. Detecting Extreme Ideologies in Shifting Landscapes: an Automatic & Context-Agnostic Approach. aug 2023.

[115] Henrique Pinto, Jussara M Almeida, and Marcos A Gonçalves. Using early view patterns to predict the popularity of youtube videos. In WSDM, pages 365–374, 2013.

[116] Swapnil Mishra, Marian-Andrei Rizoiu, and Lexing Xie. Feature driven and point process approaches for popularity prediction. In CIKM, pages 1069–1078, 2016.

[117] Mario Peucker, Thomas J Fisher, and Jacob Davey. Mainstream media use in far-right online ecosystems. Technical report, August 2022.