Responsible AI for Social Media Governance

A proposed collaborative method for studying the effects of social media recommender systems on users

November 2021

GPAI | THE GLOBAL PARTNERSHIP ON ARTIFICIAL INTELLIGENCE

Please note that this report was developed by experts of the Global Partnership on Artificial Intelligence's Committee on Responsible AI for Social Media Governance. The report reflects the personal opinions of GPAI experts and does not necessarily reflect the views of the experts' organizations, GPAI, the OECD or their respective members.

Responsible AI for Social Media Governance

A proposed collaborative method for studying the effects of social media recommender systems on users

Alistair Knott, *Department of Computer Science, University of Otago*
Kate Hannah, *Te Pūnaha Matatini, University of Auckland*
Dino Pedreschi, *Department of Computer Science, University of Pisa*
Tapabrata Chakraborti, *Big Data Institute, University of Oxford*
Sanjana Hattotuwa, *Te Pūnaha Matatini, University of Auckland*
Andrew Trotman, *Department of Computer Science, University of Otago*
Ricardo Baeza-Yates, *Institute for Experiential AI, Northeastern University*
Rituparna Roy, *Te Pūnaha Matatini, University of Auckland*
David Eyers, *Department of Computer Science, University of Otago*
Virginia Morini, *Istituto di Scienza e Tecnologie dell'Informazione, National Research Council of Italy (ISTI-CNR)*
Valentina Pansanella, *Scuola Normale Superiore, University of Pisa*

Acknowledgements

We would first like to acknowledge the other members of our project, who helped design the work reported here, and participated in project meetings. Thanks to Raja Chatila, Stuart Russell, Toshiya Jitsuzumi, Colin Gavaghan, Marta Kwiatkowska, Ivan Bratko, Koh Suat Hong, Chung Sang Hao, Przemysław Biecek, Alejandro Pisanty, Amir Banifatemi, Marc-Antoine Dilhac, Sebastian Hallensleben, Alan Paic and Jaco Du Toit. Particular thanks to Roger Taylor, who helped develop the initial project proposal. We also gratefully acknowledge the participants at our in-person and online hui (consultations). Anastasiya Kiddle and Kayli Taylor contributed to research on our community consultation project.

Several people gave valuable feedback as the project progressed. We particularly want to thank Jonathan Stray, who has been a guide throughout. Many people made useful comments on earlier drafts of the report—special thanks to James Maclaurin, Chris Meserole and John Zerilli. We also want to thank several recommender system engineers who spoke to us off the record. David Reid and Paul Ash from New Zealand's Department of the Prime Minister and Cabinet have provided essential help in our discussions with tech companies.

Ed Teather and Lama Saouma provided valuable support from CEIMIA—many thanks to you both, and to the other CEIMIA staff.

The University of Otago funded the work reported in Annex C; we are grateful for their support.

Contents

Executive summary

This report presents the findings and recommendations of the GPAI's Social Media Governance project, coordinated by its Working Group on Responsible Use of AI. Our project focuses on a question at the forefront of global discussions about social media governance: what are the *effects* of social media platforms, on individual users, and on communities? And are any of these effects harmful?

Our project considers two related questions. One question is how to *define* the concept of 'harmful' social media content. We review existing definitions of harmful content, that are used within social media companies and in wider academic and policy communities. But we also advance a particular proposal: that communities in a given country should take the lead in formulating definitions of harmful content—and that the loudest voice in these discussions should be given to the communities that suffer the most harm. Hate speech on social media is particularly directed at certain communities: in our project we have trialled a method for engaging with these groups about their lived experiences, to surface meaningful definitions of harmful content. This work focuses on a single country (Aotearoa New Zealand) as a case study, but the method we are trialling is designed to be extendable to other countries. One of our aims is to highlight the importance of regional variations in definitions of harmful content, and to propose a possible model of regional governance for Internet platforms in relation to harmful content.

A second question focuses on the AI systems that disseminate content on social media platforms, namely **recommender algorithms**. These algorithms learn about individual platform users, from their actions on the platform, and exploit this learning to deliver personalised content to user feeds. A recommender algorithm's choice of items for a given user is influenced by what it knows about this user's behaviour on the platform. But this choice also directly *influences* users' behaviour, which is in large part driven by the items arriving in their feeds. What a recommender system learns about a user thus depends in part on its earlier learning. As AI theorists have shown, this dependence makes it possible for recommender systems to push users towards arbitrary pockets of Internet content: the so-called 'filter bubble' effect. We review the theoretical models that demonstrate this effect. We also review evidence that harmful content on social media has harmful effects in the world, and evidence for a range of cognitive biases, that push social media users towards harmful content of several kinds. Taken together, these studies establish a *prima facie* cause for concern that recommender algorithms may lead social media users towards harmful content. Our project also investigates this particular concern. (This same concern has also been the focus of much recent attention, following the testimony of Facebook whistleblower Frances Haugen.)

There is an active scientific debate about what effects recommender algorithms have on platform users, and about how to measure these effects. Again we review the existing literature on these effects. Nearly all studies to date have been conducted *externally* to social media platforms, using publicly available data, obtained either through experiments on social media users and interfaces, or through tools (APIs) provided by companies to surface certain aspects of their operation. The findings of these studies are very mixed: some studies find significant harmful effects of recommender systems, some find none, and some find only small effects. We argue an important reason for this diversity is that the techniques available for studying recommender systems externally to platforms are all flawed, suffering from a range of methodological problems. In particular, none of the existing methods test properly *causal* hypotheses about the effects of recommender systems on users. These external methods simply don't provide good enough information about the effects of recommender systems on users. In particular, governments considering regulatory options for social media platforms don't yet know enough—and need to know more.

To test a causal hypothesis about the effects of a given recommender system, it is necessary to conduct experiments that *manipulate* the system, trying out different versions on different groups of users, and looking for differences in the behaviours of different groups. Crucially, this is the method social media companies use themselves, to develop and optimise their own systems. Companies are primarily interested in measuring

user engagement in their studies. The recommendation of our second project is that governments should engage with social media companies to conduct studies using these same methods to examine the effects of recommender systems on users' relationship with harmful content. This will enable them to gain a much better understanding of these effects, and provide vital information to inform subsequent policy development. It will also provide a useful new measure of transparency for social media companies. Importantly, the transparency relates to *effects* of companies' algorithms, rather than to their internal design, or the data they run on: company IP and the personal data of platform users are protected.

Again, our project focuses on Aotearoa New Zealand as a case study country. As part of the GPAI project, the New Zealand government has invited one social media company to work with us, to incorporate metrics measuring users' attitudes towards harmful content within their existing methods for optimising recommender algorithms for New Zealand users. Again, while our case study focuses on New Zealand, the proposed exercise is one that could be initiated by any government, to test the effects of any company's recommender algorithm. Note that the proposed exercise won't have any impact on user experiences: its aim is simply to learn something new from the methods companies already use for trialling their recommender algorithms, to inform policy development, and to provide a measure of transparency. Note again that the exercise provides a measure of *regional governance* of social media platforms.

Our proposed exercise involves collaborating with a social media company to study the effect of its recommender system on users' attitudes towards harmful content. The definition of 'harmful content' is once again at issue. Our practical proposal is to focus on the category of 'Terrorist and Violent Extremist Content' (TVEC), that is already the focus of productive collaboration between tech companies and governments around the world. Through the Global Internet Forum to Counter Terrorism (GIFCT), companies are collaborating in the creation and use of a shared database of TVEC material—a collaboration supported by companies and countries (including all the GPAI countries) participating in the Christchurch Call to eliminate TVEC online. Coincidentally, one of the topics for this year's Christchurch Call workstream is to explore the 'user journey [towards TVEC], and the role this may play in the broader radicalisation process'. Call participants, including all the major tech companies, have already committed to 'design a multi-stakeholder process to establish what methods can safely be used and what information is needed—without compromising trade secrets to allow stakeholders to better understand the outcomes of algorithmic processes and their potential to amplify TVEC', surfacing results between November 2021 and May 2022. This commitment to collaboration, and this timeframe, provide ideal context for the exercise we have in mind.

There are of course many legal and policy issues to address in the exercise we propose. The third part of our project, conducted by lawyers specialising in AI and social media governance, surveys these issues. Their report appears under separate cover.

Structure of the report

In **Chapter** 1 we provide definitions of key terms (recommender systems, filter bubbles, echo chambers), and survey the formal models of recommender systems that demonstrate how in principle these systems can push users towards harmful content. We also review evidence that harmful social media content can have harmful effects in the real world. Together, these reviews demonstrate a *prima facie* cause for concern, that motivates our project on recommender systems.

In **Chapter** 2 we review existing typologies and definitions of harmful online content. This review provides context for our survey of the literature on effects of recommender systems, and for our proposed fact-finding exercise: to this end, our review pays particular attention to definitions of TVEC. It also motivates our own community-focused proposal about how to define harmful content, which we describe in detail in **Annex** A.

Chapter 2 also reviews existing accounts of 'pathways towards harmful content'. One part of this review focuses on pathways to violent extremism and TVEC. Another part focuses on cognitive biases that are more generally implicated in accounts of pathways towards harmful content. These include biases towards 'moral emotional content' in political messages, towards 'moral outrage' and negative emotions, towards content focusing on 'political out-groups', towards falsehoods, and (perhaps) towards 'sensational content'. These reviews strengthen the *prima facie* cause for concern about recommender systems that motivates our project.

In **Chapter** 3 we survey existing experiments studying the effects of recommender systems on users, with a focus on experiments conducted 'externally' to social media companies, using publicly available data. We review experiments using several empirical paradigms, and find considerable diversity in their results. Our main finding is that there are methodological problems with each of the available external paradigms: in particular, they don't allow causal hypotheses about recommender system effects to be tested.

In **Chapter** 4 we survey methods for studying the effects of recommender systems used 'internally' by companies, to develop and optimise them. These methods are mainly used to examine the effects of recommender systems on user engagement, though a range of other effects are also considered.

Our key proposal is that these same methods should be used (by governments and other external stakeholders) to study the effects of recommender systems on users' attitudes towards harmful content. In **Chapter** 5 we make some specific proposals about the structure of a 'fact-finding exercise' that a government could conduct with a social media company using these methods. In the spirit of this year's Christchurch Call workstream, our focus is on studying whether a recommender system has any effect on users' relationship towards TVEC content. We suggest several possible metrics that could be used to study this question, as a starting point for discussion with companies.

We appreciate that companies already take many steps to address harmful content on their platforms, in the removal of some categories of content, and the downranking of others. We review what we know about these efforts in **Annex** B. We argue that these existing initiatives are useful—indeed, essential—but that they do not obviate the need for the fact-finding exercise we recommend in Chapter 5.

We conclude in **Annex** C (under separate cover) with a survey of the legal and policy issues that surround our proposed fact-finding exercise.

A note about the target readership for the report

Our report is written to address two somewhat distinct audiences: a 'policy' audience on the one hand (policy researchers in governments and NGOs, and public policy teams within tech companies), and an 'engineering' audience on the other hand (including tech company engineers and academic computer science / AI researchers). There is a certain amount of technical content in some places in the report, for the latter audience. To accommodate nontechnical readers, we always provide informal summaries of the key ideas, so these readers should be able to jump over any technical content and still follow the argument of the report.

1 Social media recommender systems and some issues of public debate

In this chapter, we will start by defining some key concepts for our project. In Section 1.1 we will introduce the concept of a 'recommender system' as it operates in social media platforms, and explain how recommender systems learn from the behaviour of platform users. In Section 1.2 we will outline the key concern about recommender systems that motivates our project: namely, that because recommender systems learn from platform users' behaviour, but also *influence* their subsequent behaviour, social media systems can be *unstable* in the way they curate the flow of information. In Section 1.3 we expand on this concern by giving precise definitions of two terms that are often used in discussions of social media: 'echo chambers' and 'filter bubbles'. (The focus of our report is on 'filter bubbles', as we define them, but we define both concepts, for clarity's sake. We also discuss alternative definitions of these terms, and give reasons for our choices.) In Section 1.4 we illustrate the problems that can arise from filter bubbles by describing some computational simulations of social media platforms. Our focus here is on theoretical analyses rather than empirical studies. (We will review empirical studies in Chapters 2 and 3.)

Our aim is to use these theoretical and computational analyses in Sections 1.2–1.4 to argue that there are *prima facie* grounds to be *concerned* about how recommender systems influence information flows on social media platforms—and in particular how they may encourage the proliferation of harmful content of various kinds. In Section 1.5 we emphasise the seriousness of the concern by reviewing studies showing that harmful content on social media platforms can have harmful effects in the world. We restate the *prima facie* cause for concern about recommender algorithms in Section 1.6.

1.1 What is a recommender system?

A user interacting with a social media platform finds out what's going on on the platform (and elsewhere) through a set of 'feeds', that deliver a constantly updating list of items for the user to consider. Some feeds suggest content items on the site (posts from friends or from other users); others suggest categories of content (such as Twitter hashtags); others suggest potential friends or users to follow on the site; others display adverts.

Recommender systems, in the simplest possible terms, are the machines that choose which items arrive in a given feed, and the order in which these items arrive (see Ricci *et al.*, 2015 for a general introduction). Each type of feed has its own recommender system: one system suggests content items, another system suggests potential friends, another serves ads, and so on. Each feed is organised sequentially: the recommender system that produces a feed also chooses the order in which items are presented to the user. The feeds presented to a user form the core of the user's experience on a platform; to a large degree, therefore, recommender systems are responsible for the experience of platform users.

Social media platforms are successful because recommender systems are *personalised* to users. A recommender system curates a particular stream of content items for each user, based on what it knows about this user. Personalisation happens through *learning*: a recommender system uses information about users' past interactions with content on the platform to make predictions about what they might want to see next, and the feeds they curate are conditioned on these predictions.

Of course what arrives in a user's feed is not produced by a single 'machine': it's the result of a complex mixture of algorithms and human processes. In this section, we will briefly introduce the main algorithms and processes we will be concerned with in this report. In Section 1.1.1 we'll describe the key learning mechanisms in recommender systems. In Section 1.1.2 we'll describe other mechanisms that influence what gets into user feeds, focusing on content moderation mechanisms.

1.1.1 The architecture of a modern recommender system

To introduce what a recommender system has to learn, it's useful to begin with a famous problem in recommender system design, called the 'Netflix challenge'. The challenge runs as follows: given a list of which movies a given Netflix user has watched, and a list of which movies each other Netflix user has watched, predict which movie the user will watch next. There are lots of movies, and lots of users, which adds to the challenge. This is the basic challenge for any recommender system.

One important way into the challenge is to identify general *types* of user, and general *types* of content item, so the recommender system can suggest users items of types they are known to like. There are two broad ways this classification happens. One technique, called 'content filtering', requires human experts to create at least some of the classes (see Koren *et al.*, 2009 for discussion). For instance, the Netflix recommender makes reference to the 'genre' of movies (Gomez-Uribe and Hunt, 2015). Other categories can subsequently be learned: for instance, we could learn categories of user who like particular clusters of movie genres, or we could learn to classify new movies into genres. Another technique, called 'collaborative filtering', aims to learn the relevant categories of user and content item automatically, from scratch, from analyses of user behaviour on the platform. Modern recommender systems often incorporate a mixture of content filtering and collaborative filtering methods. But collaborative filtering is particularly central to their operation.

Collaborative filtering algorithms The collaborative filtering process is often defined in relation to a two-dimensional matrix, which can be thought of as a vast table, with all the users listed down the side and every piece of content listed along the top (see again Koren *et al.*, 2009). Entries in the matrix represent user attitudes towards content items. These can take the form of 'explicit feedback' by users about content items (for instance, Netflix users' ratings of movies), or they can represent 'implicit feedback' inferred from user behaviour (for instance, what users search for, click on, hover their mouse over). In either case, the matrix will be very sparsely populated with meaningful data, because users only interact with a small fraction of available content items. A recommender system must make *predictions* about the attitudes of users towards items they haven't yet seen: the whole point of a recommender system is to suggest 'new' things for the user.

Mathematically, the process of generating these predictions can be usefully construed as a matrix factorisation task, that 'fills in the missing information' in the matrix (see again Koren *et al.*, 2009). This is a well-understood mechanism, but the matrix is far too big to perform it precisely in social media contexts.[1] A key area of research is in methods for approximating the factorisation task to cope with huge matrix sizes. An array of techniques have been deployed in this area. A summary of early methods is given by Koren *et al.* (2009). In recent years, many methods trade heavily on deep learning techniques. For instance, users and content items are often now represented as points in high-dimensional embedding spaces, that capture similarities between users and content items learned through separate mechanisms before factorisation begins (see e.g. Naumov *et al.*, 2019; Zhang *et al.*, 2021). Interactions between users and items can then be implemented as vector operations on these embedding spaces (Naumov *et al.*, 2019).

Scoring algorithms Collaborative filtering by some approximation of matrix factorisation is typically just the first operation performed by a recommender algorithm. This operation generates a shortlist of 'candidate' items to present to a given user. A second algorithm called a 'scoring algorithm' decides how these items should be ranked in the user's feed. This algorithm operates more systematically on each item in the shortlist, taking a more detailed representation of its content as input (see e.g. Covington *et al.*, 2016). The algorithm also takes a more detailed representation of the user as input, and of the user's current 'context', for instance, the user's physical location, or recent actions on the platform. Through these latter inputs, the scoring algorithm is *personalised* to the user.

As output, the algorithm produces a 'score', which traditionally expresses a prediction about how much the user

[1] The matrix used in the Netflix Challenge crossed 480,000 users with 17,770 movies, and contained 100 million ratings. In 2015, Facebook's matrix was already more than two orders of magnitude larger than this (Ilic and Kabiljo, 2015), and will of course have grown further since then.

will *engage* with this item. 'Engagement' is a complex concept. At its simplest, it might encode a prediction about how likely the user is to 'click' on the current item being processed. But modern systems measure engagement more subtly, factoring in other behaviours: for instance, how long the user views a given content item after clicking it, whether the user 'likes' an item, or shares it, or comments on it, how long the user hovers her mouse over it, and so on. (In addition, modern systems don't just learn from a history of users' behaviour *on the platform*. Today, users are tracked as they move around the Internet. A scoring system that knows, for example, that a user has recently visited a number of web sites specialising in James Bond nostalgia, and recently purchased the soundtrack to the most recent 007 movie, may on this basis give higher scores to items relating to spy movies.)

Crucially, a scoring algorithm is *trained* to produce its scores, using data about *actual* user clicks (or other user behaviours). Modern scoring algorithms are all deep networks, so the training process falls within the general field of 'deep learning'. There are two important types of training method.

Training the scoring algorithm using supervised learning The standard training method for scoring algorithms is supervised learning. In this paradigm, we present the algorithm with a set of training items we already *know* how the user engaged with (from logs of user behaviour), and we ask the algorithm to *predict* how the user will engage with these items. The difference between the algorithm's predictions about engagement and the user's actual engagement create an 'error term' that can be used to incrementally improve the algorithm, through well-known back-propagation techniques.

In practice, several different scoring algorithms are typically trained, that use different input fields and network parameters. The different algorithms are then *evaluated*, by being placed in front of different groups of users, and user behaviours are recorded. Evaluation on actual users is the acid test: the algorithm that leads most effectively to the desired user behaviours is deemed the best. ('User engagement' features strongly among the desired behaviours, because of the importance of engagement for the company's business model—though other metrics are also used, as we will describe below.) The user evaluation process will in fact be a key focus in our report: we will introduce it in detail in Chapter 4, and it is central to the project we propose in Chapter 5.

Training the scoring algorithm using reinforcement A more recent training method for scoring algorithms is reinforcement learning. Reinforcement learning methods (called 'bandit methods' in this context) incorporate the user evaluation process just described *within* the process of training the scoring algorithm. Again, we will give details in Chapter 4—but the key idea is that these methods let us search for the scoring algorithm that gets 'the best performance' when deployed on users, by modelling users' interactions with a space of possible recommender systems. Reinforcement learning methods are all about searching for the algorithm that 'performs best', on some defined metric. A key metric for scoring algorithms is high user engagement.

User engagement and other metrics We focused on user engagement in the above account of scoring algorithms. But it is not accurate to say that scoring algorithms are trained just to predict (or maximise) user engagement. Modern recommender systems are trained using engagement metrics, but also a range of other metrics, such as user satisfaction (see e.g. Zhao *et al.*, 2019) or 'meaningful social interactions' (see Stray, 2020 for a summary). Nonetheless, engagement metrics of one kind or another are still likely to be the dominant factor in training.

Summary To summarise: the basic architecture of a recommender system comprises a high-volume 'candidate generation' algorithm, that produces a shortlist of candidate content items through some approximation of matrix factorisation, followed by a lower-volume 'scoring' algorithm, that predicts for a given user, in a given context, how much that user will engage with a given item from the shortlist. These scores ultimately determine how content items are ranked in the user's feed.

We don't know the details of how commercial recommender systems are trained: these are closely guarded

company secrets. But we do know that recommender systems involve a high-volume candidate generation process, followed by a lower-volume scoring process. And we also know that both these processes focus heavily on learning *what users like to engage with*. The candidate generation algorithm learns from the full collection of user interactions with content on the platform, and identifies broad trends about which *kinds* of user engage with *which* kinds of content: it uses these trends to recommend a shortlist of items. The scoring algorithm learns to predict what *specific users* will engage with, or to produce scores that optimise the engagement of specific users. These learning processes feature metrics other than engagement—but nonetheless, what users see in their social media feeds is, to a very large extent, what the recommender system thinks they will engage with.

1.1.2 Content removal mechanisms in social media

Recommender systems aren't the only mechanisms that determine what arrives in user feeds. Some items of content *arrive* on a social media site, but are not allowed to remain there, because they are found to violate platform guidelines in some way, and have to be removed. The removal mechanisms are separate from recommender systems, but obviously they have an effect on what users see in their feeds: items which are removed are never recommended.

Mechanisms for removing content are a complex mixture of human processes and automated processes. The human processes involve the reporting of items by users, the identification of items by staff, and decisions by staff about how to define categories of removable content. The key automated processes are *content classification systems*, that operate at scale, deciding whether content items on the platform fall into any of these categories.

The presence of content removal mechanisms creates an ambiguity around the word 'recommender system'. It can be used to refer to the machine learning system that ranks items for presenting to users, as described in Section 1.1.1. It can also be used to refer to the complete process responsible for curating user feeds (as we did at the start of this section). In this report, we will use the word in the former sense, unless otherwise noted, because our main focus is on problems that can arise with the learning mechanism described in Section 1.1.1. We will now articulate these problems.

1.2 Feedback loops arising from retraining of recommender systems

A social media platform is a dynamic environment: new content is constantly arriving, and users can readily change their preferences. To keep abreast of users' changing preferences, recommender systems need to be regularly *re-trained*. This retraining process introduces an important feedback loop into the dynamics of a social media platform. A recommender system's personalised learning about a given user up to a given time-point determines the feed items that are presented to the user at that time point, using the learning mechanisms outlined in Section 1.1.1. But the user's behaviour at this timepoint *depends heavily on the items that appear in her feed*. Sometimes, the user may take her own initiatives, that are unrelated to feed items—but much of the time, she is likely to be selecting from the items that are presented to her in her feeds. If the recommender system is *updating* its learning about the user from the user's selections of feed items, then what the recommender system learns at any given time T *depends* on learning that happened at times before T, that led it to its recommendation of the feed items at T.

The feedback loop described here involves an interaction between the recommender system and each user on the platform. The recommender system personalises its recommendations to each user, by consulting a record of what this user engaged with. But the items the user engages with are often *drawn from those that the system recommended* to that user: so the system just learns to give the user 'more of the same'.

Of course, recommender system designers are aware of this problem, and take many steps to avoid it. The key intervention is to include some measure of 'exploration' in the recommender system's output for any given user. As well as recommending items it has some reason to predict the user will engage with, the algorithm should also recommend other items, about which it has no strong prediction (other than, perhaps, other

similar users clicked them). By doing so, it gives itself an opportunity to learn the user's attitude towards these items. (And it can also learn the best users for these items more generally.) Again, methods for performing this 'exploration' are company secrets—but we know some of the techniques that are involved. The 'bandit' models introduced in Section 1.1.1 provide well-known environments for exploration models, because exploration is a key component in all reinforcement learning methods: exploration techniques such as Thompson sampling, ϵ-greedy and UCB (upper confident bound) feature among the methods used (see Bouneffouf *et al.*, 2014; Zhou, 2015 for details). For our purposes, however, the key fact is that some *component* of a recommender system's output comes from its guesses about what the user wants to engage with, based on its record of the user's earlier behaviour. And this component of the system's output creates a problematic feedback loop.

1.3 Some areas of public concern: 'filter bubbles' and 'echo chambers'

The rise of online social media services has drastically changed how people interact, communicate, and access information. In traditional media, such as newspapers and television shows, content is chosen by human editors, for large groups of consumers, with consequences that are relatively simple to observe: there are many consumers, but a relatively small number of newspapers and television shows. In social media platforms, recommender systems play the role of editors, but now they select content for each user individually. This radically changes the way information flows in society. But importantly, it also makes these new information flows much harder to *observe*. This is partly because observers don't have any right to access the personalised content social media users receive in their feed. (By contrast, any observer can read a newspaper, or watch a TV programme.) But even if observers could see what users are receiving in their feeds, the fact that users each receive a personalised feed means that the new facts about information flow are massively more complex than the old ones. On top of this, the recommender systems that produce user feeds use algorithms that are private to companies—so we don't know very much about the general mechanisms that produce this complexity either.

However, what we do know in general terms about how social media systems work is sufficient to raise a number of concerns. The concerns are to do with the *dynamics* of information flow in social media systems. In particular, they relate to the possibility of positive feedback loops within these systems, that can lead to unstable effects. The two effects we will focus on are often referred to as 'echo chamber' effects and 'filter bubble' effects. These terms have been defined in several different ways, so clear definitions are important. The focus of our report is on 'filter bubble effects', but in this section, we will give technical definitions of both effects. We will also discuss other definitions that have been given, and explain why we aren't using these.

1.3.1 A definition of the 'echo chamber' effect

The concept of an 'echo chamber' is associated with a group of people who exchange opinions between themselves, and who have limited exposure to opinions outside the group. Echo chambers often feature in discussions about political opinions, since early work by Sunstein (2001; 2000). Sunstein begins by observing that the Internet has a unique capability of connecting people with similar political views. His key concern is that this Internet-facilitated exposure to those with similar opinions will lead to processes that *reinforce world views*. He worries that it will result in people hearing 'louder versions of their own preexisting commitments', and 'louder echoes of their own voices', leading to a growth in polarisation and extreme opinions (Sunstein, 2000). Note this concept of echo chambers makes no reference to recommender systems: it is purely about interpersonal communication. Sunstein's early point is simply that the Internet facilitates people getting together with others who share the same opinions.

We want to emphasise that Sunstein's main concern is with the *dynamics* of interpersonal communication (especially on the Internet). He is not concerned with groups of like-minded people communicating with one another as a fact in itself: his concern is rather with the way people's opinions might *change* as a result of this grouping. His central worry is that there may be a tendency to *amplify* existing opinions within these groups, and to progressively narrow the pool of opinions in the group. Specifically, his concern is that certain ideas within the group are more *repeatable*, and *communicable*—and that these ideas will over time come to dominate discourse within the group. Sunstein does not express this idea quantitatively, or formally, but several theorists have given more quantitative formulations of it: we will state one of these in Section 1.3.3.

1.3.2 A definition of the 'filter bubble' effect

The concept of 'filter bubbles' in social media is associated with discussions of recommender systems.[2] Again, the key idea relates to the *dynamics* of information flow on social media, and the effect that recommender systems can potentially have on this flow. The filter bubble effect is essentially that because a recommender systems learns to give each platform user 'what they want' (through the mechanisms discussed in Section 1.1.1), *the system will echo each user's own existing opinions back to them*. The basic idea is that *each user is in an echo chamber with the recommender system*, as this system is personalised to them.

Again, the key concern expressed in the concept of a filter bubble is not primarily that users will settle into some 'stable state' in their personal relationship with the recommender system. It is rather that the dynamics of a user's interaction with a recommender system can be *unstable*, and lead in certain worrying directions. Again, the key idea is that the user has certain predispositions to engage more with certain items, or certain types of content—perhaps because of existing opinions she already has, or perhaps due to general biases she has towards content with certain features. In the dynamical interaction between user and recommender system, these small predispositions can lead the user *systematically* in certain directions, or towards certain types of content. This systematic drift is 'the filter bubble effect'.

In Section 1.3.3 we will define the filter bubble and echo chamber effects more formally. But before we do, we want to emphasise three things about these concepts.

Firstly, note that the biases users have towards particular kinds of content *don't have to be large* to have large effects within a platform. The biases operate *within a dynamical system*, at every time point—the concern is that they *accumulate over time*. To take a concrete example, say a car driver is driving down a straight road. Now say her steering wheel is very slightly misaligned, taking the car to the left. At each time point, the effect of this misalignment is very small—but over time, if the misalignment persists, the car will leave the road.

Secondly, note that the 'filter bubble effect' interacts with the echo chamber effect. Very often in social media contexts, recommender systems present users with content *from their friends*. Filter bubble effects can therefore readily help to create communities of users with similar beliefs. But the core *dynamical* mechanisms in filter bubbles involve the interactions between users and recommender systems.

Finally, we want to state at the outset that not all echo chambers and filter bubbles are harmful. The Internet is full of online groups of people brought together by shared interests, passions and obsessions: these might be with a given type of music, or a given football team. (Such groups are often referred to as 'homophilous'; we will sometimes use this word too.) Groups can form amongst people facing particular challenges, such as a particular illness, or amongst people whose sexuality alienates them from their physical community. Often the existence of such groups can have incredibly beneficial consequences, for their members, and often also for wider society. But harmful material does exist, and does proliferate, on the Internet—our interest is in studying whether recommender systems have any role in this proliferation.

1.3.3 Jiang *et al.*'s formal model of echo chambers and filter bubbles

To define echo chambers and filter bubbles more formally, we turn to a model created recently by a team from Google DeepMind (Jiang *et al.*, 2019), which we believe expresses the relevant concepts particularly clearly. Readers who aren't interested in formal models can skip this section, and rely on the definitions we presented in Sections 1.3.1 and 1.3.2.

Jiang *et al.* present a formal model of a recommender system and a user, expressed as a *dynamical system*, where the state at any given time is a function of the state at the previous time. Their model is illustrated in Figure 1.1. They begin by defining a set $\mathcal{M}$ of **content items** a, that the recommender system can serve the user. They then define four time-dependent variables:

[2] The term was originally coined by Pariser (2011), in a discussion about Internet search engines. But he has also used the term in a discussion of social media recommender algorithms (see e.g. Pariser, 2015); it is now more commonly used in these latter discussions.

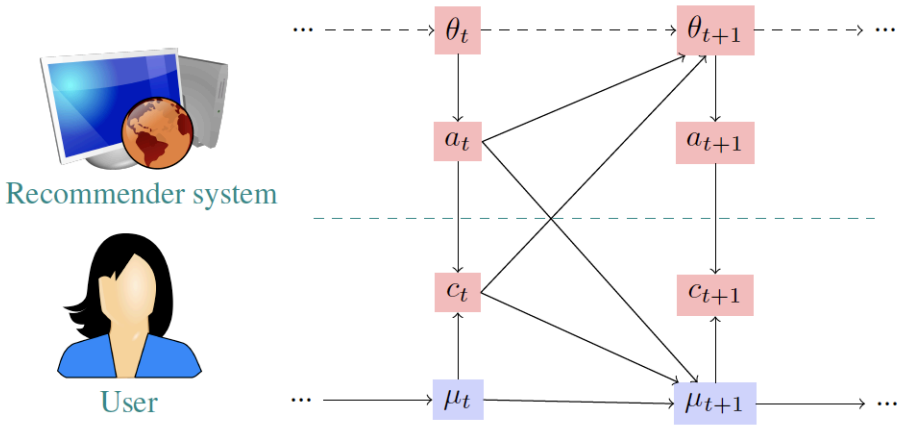

Figure 1.1: Jiang *et al.*'s (2019) dynamical system

- A **user interest function** $\mu_t : \mathcal{M} \to \mathbb{R}$, which maps each content item onto a real (positive or negative) number
- A **system recommendation** of l items: $a_t = (a_t^1, \ldots, a_t^l) \in \mathcal{M}$
- **User feedback** for recommendation in the form of a click (or no click c_t ($\{0, 1\}$ for each item in a_t)
- The recommender system's **internal model** of the user θ_t. (θ_t is an *estimate* of the user interest function μ_t.)

Jiang *et al.* then define the user interest function and recommender model dynamically. First, they stipulate that the user's interests are affected by what she clicks (and thus by what she's recommended):

$$\mu_{t+1} = f(\mu_t, a_t, c_t)$$

They then stipulate that if the user clicks on a at time t, then interest in that item increases at the next time point :

$$\mu_{t+1}(a) > \mu_t(a)$$

and if not, interest in that item decreases:

$$\mu_{t+1}(a) < \mu_t(a)$$

Finally, they stipulate that the recommender function is updated at each time point as a function of the user's clicks:

$$\theta_{t+1} = g(\theta_t, a_t, c_t)$$

Jiang *et al.* can now investigate the user's **total change in interest** from time 0 to time t. They do this by asking how the L^2 norm of this difference ($\|\mu_t - \mu_0\|_2$) changes as time advances. The L^2 norm is defined as follows:

$$\|\mu_t - \mu_0\|_2 = (\sum_{a \in \mathcal{M}} (\mu_t(a) - \mu_0(a))^2)^{1/2}$$

Jiang *et al.* analyse echo chambers and filter bubbles as conditions that lead to **degenerate effects** in this L^2 norm over time. The user's interest sequence μ_t is defined as **weakly degenerate** if the L^2 norm can get arbitrarily big:

$$\limsup_{t \to \infty} \|\mu_t - \mu_0\|_2 = \infty$$

and **strongly degenerate** if it can get arbitrarily big, and will stay big:

$$\lim_{t\to\infty} \|\mu_t - \mu_0\|_2 = \infty$$

To model a minimal **echo chamber**, Jiang *et al.* define a system where the user is constantly exposed to a particular *subset* of content items. The recommender system has no role in this process, so Jiang *et al.* set the number of recommended items (l) to 1, and hold the one recommended item constant for all t. With those settings, they prove that μ_t degenerates weakly under some (reasonable) assumptions, and strongly under other (also reasonable) assumptions.

To demonstrate a filter bubble effect, Jiang *et al.* begin by assuming that the user's interest μ has degenerative dynamics. They then examine how a recommender system *influences* this dynamics. They first consider an 'oracle' recommender system that has learned a perfect user model: that is, in their terms, where $\theta_t = \mu_t$ for all t. They then consider what happens if the recommender 'greedily' serves the user the n items estimated to be of most interest. (Recall from Section 1.1.1 that recommender systems normally incorporate some component of 'exploration': a 'greedy' system omits this component, and just serves what it thinks the user wants.) Jiang *et al.* demonstrate that this 'greedy' condition provides *the swiftest path* to degeneracy. If the recommender is less greedy, or less accurate, the path to degeneracy is slower. These latter points are shown through simulations, rather than proofs. We will survey a range of other simulation papers in Section 1.4, that make related points.

We have made this formal digression for two purposes. Firstly, it is useful to present maximally clear definitions of echo chambers and filter bubbles. To our mind, Jiang *et al.*'s definitions capture these phenomena in the right way, as conditions of instability ('degeneracy') that can arise within dynamical systems. Secondly, Jiang *et al.* prove that under assumptions that very reasonably describe what happens on social media platforms, filter bubbles and echo chambers *will arise*.

1.3.4 Alternative definitions of filter bubbles and echo chambers

There are several other possible ways of defining the concepts of filter bubbles and echo chambers. Different definitions of course lead to different conclusions about the presence of these phenomena, and of how concerned we should be about them. In this section we will outline one definition that has gained quite wide currency, namely that of Axel Bruns (see Bruns, 2019). Bruns is also concerned to give precise definitions of the phenomena in question. For Bruns, the key distinction is that echo chambers are about 'the structural properties of networks of people', while filter bubbles are about 'behavioural patterns' of groups of people. His definitions are as follows:

> An echo chamber comes into being when a group of participants *choose to preferentially connect* with each other, to the exclusion of outsiders. The more fully formed this network is (that is, the more connections are created within the group, and the more connections with outsiders are severed), the more isolated from the introduction of outside views is the group, while the views of its members are able to circulate widely within it.
>
> A filter bubble emerges when a group of participants, independent of the underlying network structures of their connections with others, *choose to preferentially communicate* with each other, to the exclusion of outsiders. The more consistently they exercise this choice, the more likely it is that participants' own views and information will circulate amongst group members, rather than any information introduced from the outside

We want to emphasise that these definitions pick out phenomena that are very different from the ones identified by our definitions. Firstly, Bruns doesn't envisage any role for recommender systems in his definition of filter bubbles. Our whole project is about problems that may arise through the use of recommender systems: in this sense, Bruns' findings about filter bubbles are simply not relevant to our project. But more importantly, we feel that Bruns's definitions just fail to articulate the key concerns that theorists like Sunstein and Pariser were expressing in their original discussions of echo chambers and filter bubbles, which relate to the *dynamics* of

information flow on the Internet (and in social media). The key worry is that these dynamics are unstable. They may well be unstable because of effects that are small at any given moment, as we discussed in Section 1.3.2—but over time they can become large, as we described informally through the metaphor of the misaligned steering wheel (from Section 1.3.2), or more formally in our presentation of Jiang's account of degeneracy.

1.4 Computational models of recommender systems

To explore the consequences of recommender systems and social media in general, a number of computational models have been developed. These abstract away from the complexities of real systems, but can nonetheless shed useful light on how social media platforms can influence the dynamics of public opinion. Many studies focus on the role these platforms can play in polarising political debates, and in radicalising groups. Despite the fact that polarisation is a natural group mechanism, it is argued by many that social media might amplify this process.

One widely employed approach is to use an agent-based model of opinion dynamics. In these models a population of agents interacts exchanging their opinions in a pair-wise or group-wise fashion. Interactions contribute to the evolution of agents' opinions, but can be modelled in different ways, to reproduce different social and technological mechanisms. The main take-home messages of these studies is that the effects of recommender systems on undesirable phenomena depend partly on social mechanisms and partly on the specific functioning of the digital platforms.

Sirbu *et al.* (2019), starting from the model of Deffuant and Weisbuch (2000), implemented the possible effects of a recommender system biasing agents' interactions. The model includes a bias parameter γ, representing the strength of personalisation: the higher the bias, the higher the chance users will interact with neighbours who are close in 'opinion space', and the lower the chance they will interact with neighbours who are distant. In this model, personalisation of the recommender system produces polarisation in situations where otherwise there would be consensus. On top of this, personalisation enhances polarisation, by pushing opinion clusters further apart in the opinion space, and by slowing down the process of convergence, compared to models with less personalisation.

Sirbu *et al.*'s simulation models opinions on a continuum. In this simulation, fragmentation arises in all settings. Other simulations model binary or discrete opinions (see e.g. Perra and Rocha, 2019; Peralta *et al.*, 2021). In these simulations, an essential ingredient for polarisation seems to be the modular network structure paired with biased (that is, personalised) selection of messages by a recommender. The rewiring of social ties in a direction tending towards homophily was shown to be another factor contributing to polarisation, alone and when considered with other polarising processes, by Rychwalska and Roszczyńska-Kurasińska (2018) in a similar context. In this study, filter bubbles and echo chambers are considered to be a consequence of individual, social and algorithmic mechanisms in a heterogeneous information environment. These effects are perhaps compounded by the limited attention and memory of users, paired with the proliferation of content: a detailed simulation by Geschke *et al.* (2019) articulates these latter effects.

To better understand how the feedback loop between users' actions and the recommender system plays a role in the dynamics of social media platforms, some simulations focus on just a single agent and their interaction with the machine learning system, disregarding network effects. Rossi *et al.* (2018) consider a news aggregator that tries to maximise the number of user clicks, and a single agent with confirmation biases. This simulation demonstrates how personalisation markedly favours the emergence of more extreme opinions, and that the strength of this emergence is related to the effectiveness of the recommender.

To summarise, computational models of recommender systems in their simplicity shed some light on how algorithmic personalisation paired with social phenomena may polarise populations, and drive individuals towards more extreme opinions in a rich information environment such as online social networks.

1.5 Evidence that online harmful content has harmful effects in the world

In this chapter, we stated our concerns with social media recommender systems in very theoretical terms. We'll summarise our concerns in Section 1.6. But before we do, we want to emphasise that problems on social media have serious consequences in the wider world. In Chapter 2 we will review many studies showing real-world effects of harmful online content of various kinds. But often these studies don't clearly *isolate* the role of harmful online content from the many other factors that cause real-world harms. In this section we will review three studies that isolate the role of online harmful content, in different ways. These studies also help to articulate our concern about recommender systems in legal terms, because they support the proposition that current recommender systems can place citizens at risk of 'actual harms' in some circumstances (as we will discuss further in Annex C).

The first study we'll mention is a criminological study of online hate speech by Williams *et al.* (2020). This study investigated whether online hate speech causes hate crimes in the real world. It measured the incidence of hate speech relating to race and religion on Twitter (geo-coded to London during an 8-month period, and also the incidence of offline racially and religiously aggravated offences in London during the same period. Crucially, it also measured 'external triggering events' such as terrorist attacks, which lead to spikes in both online and offline measures. The important finding is that the incidence of online hate speech was a predictor of offline hate crimes in London, even when these effects due to triggering events are factored out.

Müller and Schwarz (2021) used a different method to disentangle social media effects from other factors influencing off-line hate crime. Their focus was on hate crimes in Germany directed at refugees arriving in the country in 2015-2016. They looked for a relationship between the volume of anti-refugee sentiment on German Facebook and the number of refugee-directed hate crimes. Crucially, their analysis considered periods of time when Facebook suffered outages, either in local regions of Germany or across the whole country. They found that country-wide Facebook outages led to reductions in the incidence of hate crimes, and that correlations between 'refugee salience' and hate crimes disappeared in municipalities suffering local Facebook outages.

A final way of looking for causal effects is through studies that intervene in users' online experience, thus controlling for other factors. Javed and Miller (2019) conducted an experiment exposing Internet users in the US to a story about a homicide in small-town America. They varied the amount of sensationalist language that was used, and whether the suspect was identified as a Muslim (that is, as a member of an 'out-group' for the respondents). They found sensationalist language increased users' support for violent retribution, while identification of the suspect as a Muslim increased support for anti-Muslim policies.

1.6 Summary: A prima facie cause for concern about recommender algorithms

In this chapter, we introduced social media recommender systems and explained how they learn (Section 1.1), and discussed some reasons for concern about the effects they can have on currents of social media opinion (Sections 1.2). We spelled these out in Section 1.3, in a formal model of a recommender system, and a formal demonstration of instabilities due to 'filter bubbles'. In Section 1.4 we elaborate these concerns by showing effects that arise in computer simulations of recommender systems, indicating how these systems can contribute to polarisation and proliferation of harmful content. In Section 1.5 we reviewed evidence that harmful online content can lead to harmful actions in the real world, which emphasises that real harms are at issue.

We don't want to draw any firm conclusions about recommender systems from the analyses and data presented in this chapter. Our argument is just that the analyses and findings presented in this chapter give us legitimate reasons to be *concerned* about the effects of recommender systems on the dynamics of online opinion. If there are reasons for concern, then we need to look more closely at how social media platforms work in the real world. That is what we will do in the remainder of this report.

Before moving on, we will step back for one moment to ask *who* should be concerned about the effects social

media platforms may have on public opinion dynamics. Who has skin in this game? One possible answer is 'the general public'. But in practical policy-making terms, it is useful to identify 'stakeholders' of various kinds, with an interest in social impacts of AI technologies—for instance, governments and civil society groups. Within this set of stakeholders, we should include social media companies themselves: social media companies obviously have a strong interest in ensuring that their products don't have harmful effects on their users. We will return to the question of identifying stakeholders, and working out how they can collaborate, in Chapter 5.

2 Definitions of 'harmful content': destinations, pathways and classifiers

The key concern we have just raised is that recommender systems may have a tendency to direct users towards *harmful* content of various kinds. Before reviewing empirical work exploring this idea, in this chapter we will discuss the concept of 'harmful content', which is inherently complex and must be approached carefully. The concept of a 'journey towards harmful content' is also complex and challenging, and we will discuss this concept in the current chapter too.

We will begin in Section 2.1 by discussing methodologies for identifying harmful content, focusing on the methodologies we are employing in our own project, and acknowledging some important limitations on our ability to discuss harms that are directed towards groups that we (the report authors) do not belong to. In Section 2.2 we will review some commonly discussed categories of harmful content, with particular attention to Terrorist and Violent Extremist Content (TVEC), which is a focus for the current report. We then review existing accounts of 'pathways towards harmful content', focusing on TVEC specifically in Section 2.3, and discussing more general factors in Section 2.4, with a summary in Section 2.5. We conclude in Section 2.6 by reviewing methods for classifying online content into the kinds of category discussed in Section 2.2.

2.1 Methodologies for identifying harmful content

Harmful content is often directed towards particular groups in our society. There is often a focus on groups that are already vulnerable or disadvantaged. The community consultation part of our project discussed in Annex A centres around the principle that these groups should have the strongest voice in definitions of harmful content. In that project, we are actively trialling a methodology for defining categories of harmful content that are relevant to a particular country (Aotearoa New Zealand). This methodology acknowledges that definitions of harmful content often need to vary from one country to another. It proposes that representatives from the groups that are subject to the most harm online should be most heard in discussions about how harm should be defined.

While this work is under way, there is already some consensus about broad categories of harmful online content. And there are certain categories of harmful content where the main harm is to the groups that produce or consume the content, rather than to an out-group. (Antivaxx content is a current case in point.) In this chapter, we will review the broad categories referenced in current usage, identifying areas of difficulty where these arise, and we will discuss ideas about the origins of these categories of harmful material. But we want to note that many of the authors of this report come from privileged groups, and do not have lived experience of the harms being discussed here.

2.2 Commonly identified categories of 'harmful online content'

In this section, we will review some widely accepted categories of 'harmful content'. There are not many broad taxonomies of harmful content; most studies focus on specific categories of harm. But a useful recent taxonomy is given by Banko *et al.* (2020), which we reproduce in Figure 2.1. From this taxonomy, we will discuss extremism/terrorism (Section 2.2.1), illegal sexual and violent material (Section 2.2.2)[1], hate and harassment (Section 2.2.3), misinformation (Section 2.2.4), self-inflicted harm (Section 2.2.5), and a category

[1]We won't include 'adult sexual services' in our review; this category is less obviously harmful than the others in Banko *et al.*'s diagram.

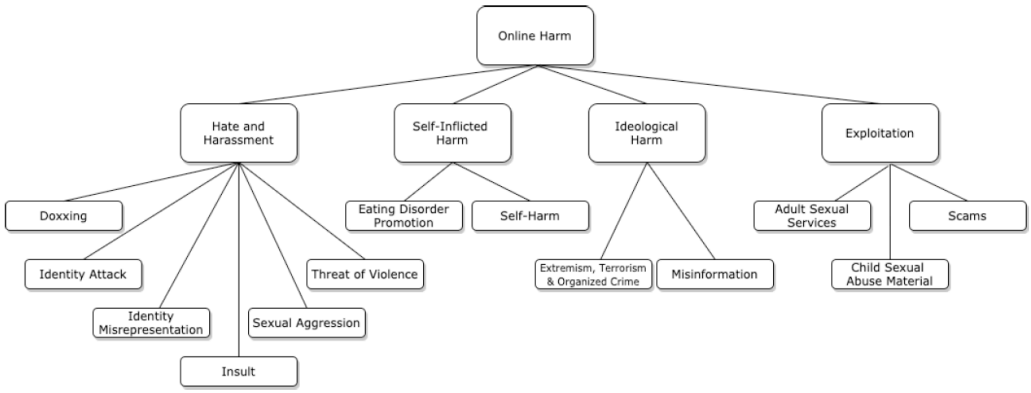

Figure 2.1: Banko *et al.*'s (2020) taxonomy of harmful content

cutting across several classes, conspiracy theories (Section 2.2.6).[2] We'll discuss methods for training classifiers to identify material from these categories in Section 2.6.

2.2.1 Terrorist and violent extremist content (TVEC)

One clear example of harmful content is 'terrorist and violent extremist content' (TVEC). This concept is closely associated with the work of the Global Internet Forum to Counter Terrorism (GIFCT). GIFCT was established in 2017, as a collaboration between Facebook, Microsoft, Twitter, and YouTube, and now includes 17 tech companies. Its mission is to 'prevent terrorists and violent extremists from exploiting digital platforms'. GIFCT and its mission gained considerable prominence following the Christchurch Call initiative, that took place in May 2019 after the Christchurch mosque attacks in March that year. The Christchurch Call brought together governments as well as companies, in a commitment 'to eliminate terrorist and violent extremist content online'. GIFCT can be understood as implementing some of the commitments of the Christchurch Call: in its own words, it 'institutionalises the spirit of the Call' (GIFCT 2021b). GIFCT is governed by a Board formed from representatives from the participating companies, but it is also guided by an Advisory Committee made up of representatives from civil society, government, and intergovernmental organisations.

GIFCT doesn't have an official definition of TVEC. Indeed only some of its member companies have detailed public definitions of terrorist and violent extremist material, and these definitions do not perfectly align (OECD, 2020). However, GIFCT defines several concrete measures relating to TVEC, that amount to a practical definition of sorts. Most importantly, the organisation curates a 'hash database' of content items that are agreed to be TVEC, that are shared between companies, and defines a protocol surrounding the use of this database. (Hashes are unique digital fingerprints associated with content items, which makes them easy to identify and search for.) There are still many unresolved issues around how to define TVEC content (see GIFCT, 2021a for a nuanced discussion). At present, GIFCT uses a mixture of 'list-based' methods (focusing on identified organisations and individuals) and 'behaviour-based' methods (focusing on content) to decide about the contents of the database. But there is a clear operational principle around use of the database: any item of content that is added to the database, and thus qualifies as TVEC, is deemed to be harmful enough that participating companies should take it down immediately.

Originally, GIFCT's hash database contained mainly videos and images, and focused on Islamic terrorist organisations. (Categories included 'graphic violence against defenceless people', 'imminent credible threats', 'glorification of terrorist acts' and 'recruitment and instruction', along with content relating to specific incidents.) But earlier this year, GIFCT proposed various extensions to the categories of material shared in the hashing database, to include written manifestos, and to include far-right and white supremacist groups

[2]We won't consider scams or bullying in our review, because recommender systems are less implicated in the propagation of these types of content.

more systematically (see GIFCT, 2021a). As a concomitant of these expansions, GIFCT proposed increased transparency and auditing around decisions about hashing, along with a company-internal appeals process, and various 'newsworthy, academic and legal exceptions'.

A number of countries have enacted laws about TVEC-related online material, defining particular categories of content that are illegal, and must be removed by companies in their jurisdictions. As with companies, these definitions vary from country to country: an excellent recent review of the differences is given in OECD (2020). But the GIFCT's hash database still appears to be the key 'executive' mechanism in defining, and responding to, TVEC content worldwide. (Companies may well not have uniform access to this database, however, as we discuss in Annex C.)

2.2.2 Harmful sexual and violent material

Criminal law also identifies more general categories of illegal content. In all countries, underage sex is illegal, and images of underage sex are also illegal (though definitions of 'underage' vary between countries); accessing sexual content is also normally restricted to those above a certain age. Content that promotes, incites or instructs in violence is also illegal in most jurisdictions. But there are more differences between jurisdictions here. Social media companies also have their own operational policies on the kinds of sexual and violent content they remove (see Annex B for further references). Sexual content is typically removed by default, to limit its use in harassment, as well as to ensure compliance with underage content laws. Violent and 'graphic' content is typically also removed by default. Some instances of sexual and graphic content are exempted from these bans, if there are good educational or awareness-raising grounds for presenting them, and companies all have mechanisms allowing users to query bans, and to reinstate content if there are good grounds for presenting it.

Our study of recommender algorithms won't focus on violent/graphic content, outside the case of terrorist and violent extremist activities. It won't focus on content relating to criminal activities either. There are certainly incremental 'pathways to criminality', and there is some evidence that social media shapes criminals' social identities (see e.g. Boduszek *et al.*, 2021), but we choose not to explore the role of recommender systems in this process.

2.2.3 Hate speech

A broad and complex form of harmful content is hate speech: hateful material directed towards a particular group of people. Hate speech has been defined in several ways. As summarised by Sellars (2016), some definitions emphasise the *intentions* behind hate speech (which are most relevant for criminal sanctions); others emphasise its *perception* by recipients; others emphasise its *content*. Our community consultation project draws on all of these perspectives, as we will discuss in Annex A. As a methodological principle it pays special attention to recipients as informants about definitions of hate speech. But the consultation process still results in examples of content classified in particular categories, which pay attention to both intentions and content.

Hate speech is often defined as being directed towards a *group* of people, rather than towards an individual. Hate towards individuals is often better classed as 'bullying' or 'abuse'. These topics won't be a focus of our current study. But hate towards individuals who *represent* a particular group (such as politicians) is certainly in scope for us. A taxonomy of hate speech targets defined by Kiritchenko and Nejadgholi (2020) usefully distinguishes between 'people' and 'entities', with subcategories of group-directed hate speech identified in each category; we reproduce this taxonomy in Figure 2.2.

Antagonism towards a certain group often emerges from a situation of *polarisation*, where two groups establish themselves in opposition to one another. (We will discuss this process more in Sections 2.3 and 2.4.3.) As polarisation becomes extreme, hate speech can arise symmetrically, from each group towards the other. But hate speech can also arise unilaterally, from one group to another group which doesn't reciprocate these hateful feelings. Often, marginalised groups aren't in a position to reciprocate hate directed towards them—or at least, not publicly. We are interested in hate speech both in 'symmetrical' and 'unilateral' contexts.

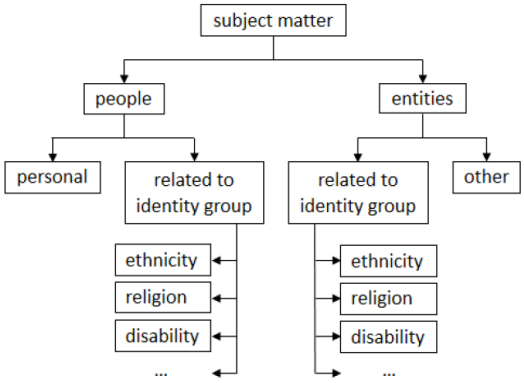

Figure 2.2: Kiritchenko and Nejadgholi's (2020) taxonomy of hate speech targets

While some studies distinguish discretely between 'hate speech' and 'non-hate speech' for practical purposes (see e.g. Lingiardi *et al.*, 2020), there is a consensus that hate speech exists on a continuum of intensities. Identifying hate speech at all degrees of intensity is of considerable importance, if we are to track its earliest manifestations, and to understand the mechanisms through which hate becomes more intense. Several methods have been proposed for classifying the intensity of hate speech. Some proposals put forward a graded set of specific categories; for instance Bahador (2020) proposes a series of categories running from 'disagreement' through 'negative actions', 'negative character' to 'demonising and dehumanising' to 'violence' and 'death'. A number of recent papers (Aroyo *et al.*, 2019; Kritchenko and Nejadgholi, 2020) propose annotating hate speech on a continuous scale of severity, and additionally suggest a method for performing this annotation, based on comparative rather than absolute judgements, which has been validated in other domains (Goffin and Olson, 2011). Continuous scales are particularly useful in resolving disagreements between annotators, because average ratings can readily be taken.

Other proposals use structured annotation schemes, to capture different dimensions of intensity. For instance the scheme of Paasch-Colberg *et al.* (2021) annotates predicates and arguments independently, and then empirically studies how these are combined. The schemes of Salminen *et al.* (2018) and Vidgen *et al.* (2019) similarly distinguish types of hateful language (accusations, humiliation, swearing, promoting violence) from targets of hate. The scheme of Waseem *et al.* (2017) distinguishes between 'implicit' and 'explicit' articulations of hate. A useful summary of recent schemes is given by Kiritchenko *et al.* (2021).

Some research focuses on particular stages in the development of hate towards a group. A stage often singled out for attention is the 'othering' of people from the target group—a process whereby they are construed as less worthy of consideration in some way, or even as less than fully human (see e.g. Zickmund, 1997; Kteily *et al.*, 2015). We will discuss this process more in Sections 2.3 and 2.4.3.

The targets of hate speech are diverse, and often very country-dependent, as we emphasise in our community consultation project (Annex A). It is still useful to single out a few broad categories, though. One concerning trend is for opposing *political* communities to start expressing hate towards one another: that is, for groups formed around alternative political opinions start conceiving of the people in other groups as objects of hate. This kind of 'political sectarianism' has been on the rise in the US for many years, at least since the 1980s (see Finkel *et al.*, 2020 for a recent review). Another common target for hate speech is women. While this trend is endemic in human cultures, there are particularly concerning trends in online spaces, with the rise of 'incel' culture: incel content is increasing rapidly online (see e.g. Hoffman *et al.*, 2020; Papadamou *et al.*, 2021). Other common targets for hate speech include foreigners, especially refugees (see e.g. Müller and Schwarz, 2021), people from particular religious groups (see e.g. Zannettou *et al.*, 2020), and people from LGBTQ+ communities (see e.g. GLAAD, 2021). A more detailed quantitative analysis of the targets of hate speech on social media is given by Mondal *et al.* (2017).

2.2.4 Misinformation and disinformation

Misinformation is an umbrella term, covering information that is false, inaccurate, or misleading. The term applies both when the purveyor of the information is aware of its dubious status, and when the purveyor is unaware. In the former case, we can also use the term 'disinformation'. The communication of disinformation is a strategy used by extremists, among others, to win adherents to a cause. (Disinformation can also be used in various forms of fraud and hoax, but we will focus on its use by extremists, in propaganda for a given cause.)

In general, misinformation is harmful, because people (and communities) should be guided in their actions by true facts. But misinformation often arises because of mechanisms in specific domains. The two principal domains for misinformation at present are medicine and politics. We will briefly review each of these domains. A broader review is given by Van Bavel *et al.* (2020).

2.2.4.1 Medical misinformation

Medical misinformation can be defined as information that runs explicitly contrary to 'mainstream medicine'. Of course there are many debates and disagreements within mainstream medicine; misinformation is best defined as information whose incorrectness is a matter of broad consensus within the medical community.

There are several common varieties of medical misinformation. One common type is misinformation about vaccines and their effectiveness. Skepticism about vaccines in general predates the current skepticism about the Covid vaccine, though of course it helped create the conditions in which this latter skepticism flourished. The current skepticism about the Covid vaccine is linked to skepticism about the severity of Covid-19, and to various unwarranted ideas about its medical nature, and about ways of treating it (Smith and Graham, 2019), which can all be regarded as types of medical misinformation. There are several proposals about how and why medical misinformation is propagated online—but the placebo effect (Evans, 2003) is likely to be a central mechanism: people who believe a treatment they are using is effective gain benefit from it, even if it has no physical effect. Placebo effects, coupled with people's tendency to infer causal effects from correlations in their environments (Boudry *et al.*, 2015) are a powerful mechanism for creating and propagating medical hypotheses outside mainstream medicine. Another powerful mechanism is the desire of sick patients to be cured, which leads to a very understandable openness towards hypotheses (Martin, 2008). These mechanisms have particular traction in 'alternative health' and 'wellness' spaces in social media, where there is already a degree of skepticism towards mainstream science and medicine (see e.g. Johnson *et al.*, 2020).

Many types of medical misinformation are harmful—but the harms of misinformation about the current Covid epidemic have been particularly well documented. For instance, Loomba *et al.* (2021) showed in a randomised controlled trial of US and UK groups that currently circulating misinformation about the Covid vaccine induced a drop of over 6% in people reporting they would 'definitely take' the vaccine. Given the effectiveness of Covid vaccines in reducing hospitalisation and mortality (see e.g. Haas *et al.*, 2021), misinformation about Covid vaccines has a measurable harmful impact.

2.2.4.2 Political misinformation

Political misinformation can be defined as demonstrably false statements about politicians and political groups, and false or implausible attributions of motives to politicians or political groups. Much of what goes under the term 'fake news' is political misinformation—though notoriously the term is also used by purveyors of political misinformation, labelling true statements as false.

Political misinformation can be harmful, in that it can impair the functioning of democracy: if large numbers of voters believe false propositions, they can sway the results of elections in harmful directions (see Persily and Tucker, 2020 for some recent case studies). The harm caused by misinformation in these cases is, of course, hard to quantify. But there have been careful analyses of several recent elections where misinformation on social media appeared to play an important role. For instance, in the 2016 Philippines election, large amounts of misinformation appeared on social media to support Rodrigo Duterte's campaign (see e.g. Ressa,

2016). In a retrospective analysis of Facebook, Sinpeng *et al.* (2020) found that Duterte's posts were far more 'viral' (shared and commented on) than those of his rivals; that Duterte's supporters were more likely to post negatively about other candidates, and more likely to rate Facebook as 'trustworthy' than supporters of other candidates. A large amount of misinformation also circulated on social media during the 2016 US election campaign. According to Silverman (2016), the most popular fake election news stories generated more total engagement from US citizens on Facebook than the most popular election stories from 19 major news outlets. A more quantitative analysis of web traffic data and surveys (Guess *et al.*, 2020) shows that Facebook had a particular role in spreading content from 'untrustworthy' websites. Guess *et al.* also show that consumption of news from these untrustworthy sites came primarily from the 20% of Americans measured as having the most conservative information diets. They describe the echo chamber for political misinformation during the 2016 election as being 'deep', but 'narrow' (though its narrowness might be disputed, if it extends to 20% of the US population).

Misinformation on social media has also served to entrench regimes established by political coups. Here again, amplification on social media is recognised as having played an important role. The 2021 military coup in Myanmar is a particularly clear case, where Facebook had a role in amplifying incitement to violence as well as political misinformation (see e.g. Global Witness, 2021).

2.2.5 Harm-advocating content

A pernicious category of harmful content on social media is content advocating self-harm of various kinds. The harm can involve extreme dieting, self-mutilation, or even suicide.

Eating disorders are associated with social media use in complex ways (see Padín *et al.*, 2021 for a recent review). This general association is likely due to social media's proliferation of powerful social norms about beauty. For young women, consuming social media can lead to depression, and a cycle of increasing consumption, as revealed in Frances Haugen's recent disclosures about Instagram. But there are also even more concerning trends on social media towards positive advocacy of eating disorders, or 'pro-anorexia' stances.

Pro-anorexia groupings likely originate from communities sharing advice about dieting. But within these communities, certain behaviours appear to be reinforced: for instance, there is some evidence that pro-anorexia content is more frequently 'liked' by users than informative anorexia content (Syed *et al.*, 2013). Pro-anorexia groupings also create norms around anorexia (see e.g. Tierney, 2008). In these groupings, negative disclosures touching on 'stigma-related' aspects of anorexia can be reinforced (Chang *et al.*, 2016). These tendencies can push dieting communities on social media towards dangerous practices.

Communities focused on self-harm also exist on social media. These communities are environments where people who self-harm can support one another (Lavis and Winter, 2020), and interact openly about practices that are taboo in wider society. But again there is evidence that certain extreme behaviours are reinforced within these communities; for instance, in a study of Instagram, posts that mentioned an intent to self-harm or to attempt suicide received more likes than those that did not (Carlyle *et al.*, 2018). Again, these tendencies can have the effect of normalising self-harm content within online groups, and of pushing people in these groups towards increasingly dangerous practices.

There is good evidence that online exposure to self-harm content is predictive of actual self-harm and suicide-related behaviours For instance, Arendt *et al.* (2019) show through a longitudinal study that exposure to self-harm content on social media predicts susceptibility to suicide and self-harm outcomes one month later. The authors used a two-wave survey among 729 US young adults, with a gap of one month between waves. The participants were exposed to harmful material in both waves, but inclination towards self harm was measured only in the second wave. The social media platform used as case study was Instagram. It was found that 20.1% of the candidates actually searched for harmful content by themselves. But 63.9% said that the exposure to harmful content did emotionally disturb them.

Harm-advocating content is unusual in the typology of harmful content in respect of who is harmed. For categories like hate speech and TVEC, the harmed community is emphatically an out-group, separate from

those who produce and engage with the material. But for harm-advocating content, the harmed group is precisely those who produce, or engage with, the harmful material. Medical misinformation arguably patterns with harm-advocating content in this contrast: those who produce or engage with it are often the ones who are most harmed. Political misinformation is directed towards particular groups, who are the immediate targets of harm—but more broadly, it can also be thought of as harming the democratic system as a whole, and therefore the whole community.

2.2.6 Conspiracy theories

Conspiracy theories assert a bewildering array of false facts. The mechanisms which select extreme practices in self-harm communities seem to operate in conspiracy communities to select beliefs on a range of other principles, which are usefully explored by Boudry *et al.* (2015). Key principles are 'promiscuous teleology', a predisposition to seek explanations for events in intentional behaviours of other individuals (see e.g. Shermer, 2011), and a tendency to identify causal patterns where none exist (which may have been an adaptive bias during human evolution; see Haselton *et al.* 2015), along with the 'placebo thinking' already mentioned in Section 2.2.4.1.

In one sense, the material espoused by conspiracy theorists could be classed as 'misinformation' (already discussed in Section 2.2.4)—but the material in conspiracy theories exhibits another interesting tendency, which is for different varieties of harmful content to gradually *coalesce* online. This tendency can be seen in the current Covid epidemic, in the gradual forming of links between anti-vaccine groups and a diverse range of other groups. Compelling evidence for these links is shown in Johnson *et al.*'s recent study of anti-vaxx communities on Facebook (Johnson *et al.*, 2020). Anti-vaccine groups are blending with anti-capitalist groups (with Bill Gates as a focus), with 'Christian Nationalist' groups (Whitehead and Perry, 2020), and with extreme political groups on both the left and right (Sutton and Douglas, 2020), as well as with wellness groups (as discussed in Section 2.2.4.1). Often these blends are articulated through the assertion of extremely bizarre facts: for instance, the suggestion that Bill Gates is inserting microchips into Covid vaccines, which was believed by 44% of US Republicans early in the pandemic (Romano, 2020), or the assertion made by QAnon supporters that Satanist paedophiles congregated in a pizzeria in Washington, DC. The references to paedophilia and Satanism in QAnon highlight rapprochement with another category of harmful content—the taboo sexual/violent content discussed in Section 2.2.2.

2.3 Pathways towards harmful content: a focus on TVEC

The main focus in our GPAI project is in the mechanisms that draw people towards harmful content. We want to study whether social media algorithms play a role in these mechanisms. In the next two sections, we will review what is known about these mechanisms, and what role (if any) social media may play in them.

In the current section, we will focus on mechanisms that draw people towards TVEC content, because this is the focus of the 2021 Christchurch Call workstream, and of our practical research proposal. In Section 2.4, we will review some broader cognitive biases, that could have a more general role in drawing people towards harmful content. But before we begin, we want to state emphatically that the 'early stages' of the pathway to TVEC are not remotely problematic: they are entirely normal. Our reason for even considering the early stages of the pathway is that the mechanisms that move people from one point in the pathway to another may be the same at different points on the pathway, as we will discuss in more detail in Section 2.3.5.

The processes that lead people to violent ideological positions are the subject of a vast and growing literature. Two broad perspectives are taken in this literature: one focuses on the psychology of individuals, and one focuses on the behaviour of groups. The interactions between these two perspectives are important, because becoming a violent extremist typically involves strong psychological *identification* with an extremist group. We will begin by reviewing an oft-cited and relatively theory-neutral account of the psychological journey towards extremism, given by Moghaddam (2005), and then re-examine this journey from various more theoretical and more empirical perspectives.

2.3.1 Moghaddam's (2005) analysis

Moghaddam views the journey towards extremism through the metaphor of a 'narrowing staircase', leading to successively higher floors of a building. At each floor, only a minority of people move to the next floor, so the staircase also acts as a filter. At each floor, people's perceptions of the building also (and their journey within it) also change: in particular, their perception of the *choices* available to them progressively narrow, until the only choice they perceive is violence. The ground floor is occupied by large populations, subject to a variety of real or perceived injustices and deprivations. Movement to the first floor happens particularly for those individuals who *perceive* injustices, whether they are real or not—and particularly for individuals who see injustices as applying *collectively* to some group they belong to, rather than just to themselves as individuals.

On the first floor, people are motivated to find ways to *address* the injustices they perceived. They search for different options: some options (such as participation in democratic processes or pursuing legal redress) take them out of the building altogether. The staircase to the second floor is for those who don't find these options, but are still motivated to act.

On the second floor, the desire to act and the (real or perceived) paucity of options coalesce in the direction of aggression towards some enemy or out-group. Those with sufficient aggression climb to the third floor, where the first organised activities of extremist groups are encountered. On this floor, a process of 'moral re-engagement' begins, in which individuals develop a new moral code, inconsistent with mainstream morality. This happens through various explicit tactics—in particular practices involving 'isolation' (from the mainstream community), 'affiliation' (with the extremist group), 'secrecy' (re-engagement must happen out of the public eye) and 'fear' (of government measures against the group). Those on whom these tactics are effective gain access to the fourth floor, where people are full members of an extremist organisation.

Options for people on the fourth floor are very limited. For Moghaddam, whose analysis focuses on Islamic extremism, there is 'little or no opportunity to exit [this floor] alive'—but even for other forms of extremism the option of violence looms increasingly large. On this floor, people fall under the authority of the leaders of the extremist organisation, and are strongly constrained by this authority as well as by their exclusion from mainstream society. There are few options other than to prepare psychologically for acts of violence. This preparation requires people to distance themselves from those who will suffer from the violent act, and to rehearse 'terrorist myths' articulating the benefits that will come from it. Those who complete this preparation move to the fifth floor, where actual violence is perpetrated.

2.3.2 Social identity theory as a framework for analysing paths to extremism

The processes described by Moghaddam (2005) can be construed in useful ways from various theoretical standpoints. A particularly helpful standpoint is social identity theory, which was developed as a theory of intergroup conflict (Tajfel and Turner, 1979), and has often been used to describe terrorist individuals and communities since then (see e.g. Schwartz *et al.*, 2009; van Stekelenburg, 2014). These descriptions often trade on a useful distinction made by Klandermans and De Weerd (2000) between 'social identity' and 'collective identity'.

A 'social identity' is a cognitive entity, inside the mind of an individual, that conceptualises this individual (to herself) as a member of some given social group. A person can have many social identities, which become salient for the individual in different social contexts.[3]

A 'collective identity', on the other hand, is a social phenomenon, involving the formation of an actual group in society. The formation of an actual social group involves *actions* by the members of the group. (Klandermans and De Weerd construe it as 'a process that is constantly under way', rather than an established fact, to emphasise that social groups require active maintenance in order to persist.)

Two mechanisms link social identity and collective identity. One involves a perceptual process, whereby an

[3]She also has a 'personal identity', distinguishing her from other people, but this doesn't feature in accounts of radicalisation.

individual activates a given social identity through classifying herself as a member of a perceived social group, and as distinct from other groups. This is called 'social categorisation'. The other involves action—specifically, the process of acting in accordance with an adopted social identity. This is called 'social identification'. Both processes are crucial in connecting social identity (an individual's conceptualisation of group membership) with collective identity (observable social groups maintained through overt actions). Both processes can can also be seen as central in an individual's pathway towards violent extremism. van Stekelenburg (2014) develops this idea explicitly in her account of the journey towards extremism.

2.3.3 van Stekelenburg's (2014) analysis

van Stekelenburg sees the journey towards extremism as involving a roughly sequential process of 'politicisation', then 'polarisation', then 'radicalisation'. This analysis is a fairly standard one, which has parallels in Moghaddam's account; but van Stekelenburg emphasises the 'tough identity work' that is required by individuals at all stages of the journey.

Politicisation involves the identification of social groupings in the world (a process of social categorisation), and the association of oneself with one particular group (the activation of a social identity). To count as politicisation, one's own group must be perceived as experiencing some form of injustice in relation to other groups. This parsing of the social world into groups receiving unequal treatment is an early (and in itself entirely harmless) step in the journey towards extremism.

Polarisation involves emphasising membership of one's own group, and defining another group as an 'out-group'. This happens through action (social identification) as well as through perception. People treat members of their own group favourably compared to members of an out-group in many experimental paradigms (see Everett *et al.*, 2015 for a review). The factors that influence this discrimination are still being explored, as are the effects discrimination has on participants (see e.g. Aberson *et al.*, 2000; Hunter *et al.*, 2019). The key finding is that polarising actions reinforce boundaries between groups in the world, as well as between groups as conceptualised by individuals: categorisation and action processes can readily reinforce one other, and exacerbate polarisation. Finkel *et al.* (2020) argue the political sectarianism developing in the US between Democrats and Republicans is an example of this process. Their analysis is particularly interesting in distinguishing 'out-party hate' from 'in-party love' (factors which are often confounded in experimental designs). They show that out-party hate has become a stronger factor than in-party love in the last ten years or so. More generally, the polarisation process tends to involve the production of hate speech, of the kind discussed in Section 2.2.3. As polarisation proceeds, the amount and intensity of hate speech towards an outgroup grows; graded categories of hate speech, such as those proposed by Bahador (2020), are useful for charting this process.

If polarisation becomes strong enough, radicalisation can develop. A 'radical' (in van Stekelenburg's terminology) is someone who is prepared to commit violence for political ends. Again, radicalisation is seen as involving intertwined processes of conceptualisation of groups (by individuals) and active formation of groups (by collectives of individuals). The groups in question are now radical *subgroups* of the original groups, whose members are distinguished from 'moderate' group members as well as from the original out-group. The actions through which individuals identify themselves as a member of a radical subgroup include the 'moral re-engagement' practices associated with Moghaddam's third floor, that strengthen conceptualisations of the radical group.

2.3.4 Further theoretical analyses of the radicalisation pathway

The basic account of the radicalisation journey sketched above has been added to by many theorists. We will briefly review some important extensions.

Neumann (2003) notes a disagreement among theorists as to whether the endpoint of radicalisation is defined primarily in cognitive terms, as a certain *mindset*, or set of ideas, or in terms of actions, or dispositions to act. Some commentators want to emphasise action-based definitions of terrorists (see e.g. Horgan, 2011), while others emphasise cognitive definitions (see e.g. Smith, 2009). Neumann argues that cognitive and action-related mechanisms are intimately interconnected, and should not be separated; this is also clearly the position taken by van Stekelenburg (2014), and we support this position too. Neumann notes the definition

used by the US Department of Homeland Security squarely combines cognitive and action-related elements (see Allen, 2007): radicalisation is defined here at 'the process of adopting an extremist belief system, including the willingness to use, support, or facilitate violence'.

McCauley and Moskalenko (2008) emphasise the 'reactive' nature of many of the transitions that take place during radicalisation, drawing on a diverse range of actual cases. They see the step of joining a radical group as involving individual autonomy, but see a strong role for 'external circumstances' in many of the other steps. There appears to be particularly good agreement that the start of radicalisation is initiated by a 'perception of personal crisis', that produces a 'cognitive opening' for radical ideas (see Baugut and Neumann, 2020 for a review).

Several theorists emphasise the role of interpersonal networks in the radicalisation process (see again Baugut and Neumann, 2020 for a review). This emphasis is consistent with van Stekelenburg's account, in which social groupings are also central. Some accounts pay particular attention to hierarchical structures in radicalising groups; see in particular the account of Gendron (2017).

Several theorists foresee a particular role for 'moral disengagement' in the journey to violent extremism (see Frissen, 2021 for a summary of these analyses, and some empirical support). Moral disengagement is related to Moghaddam's process of 'moral re-engagement', but Frissen appears to see the process operating at several points during radicalisation, rather than at one particular point.

It is also very important to include analyses of trajectories towards right-wing / white extremism in the broad account given above. Moghaddam's account arguably focuses on Islamic terrorism, but van Stekelenburg's account is intended to cover right-wing extremism too. In relation to the right-wing radicalisation pathway, it is worth noting that analyses from within this movement often speak of 'red-pill' moments of particular significance (see e.g. Tait, 2017). But again there is no specific point in the radicalisation pathway identified by this terminology: as discussed by Munn (2021), it is used to describe several different points in the pathway.

To conclude this summary, we note a fairly strong consensus among theorists that extremists are not crazy or pathological (see Moghaddam, 2005; Post, 2007), and don't fall into neat demographic or economic groups (see again Moghaddam, 2005; Rae, 2012). Particular roles in a terrorist organisation may be associated with demographic profiles, however, as discussed by Perlinger *et al.* (2016). There is also a fairly broad consensus that terrorists are not characterised by particular personality types (see again Rae, 2012), though there is some indication that those with high levels of anger and anxiety have a higher predisposition to support extremist groups (Anastasio *et al.*, 2021).

2.3.5 Are the early stages of the radicalisation pathway relevant?

We want to conclude our review by reiterating that none of the theorists we are discussing here view the *early stages* of the radicalisation journey as being problematic in themselves. Moghaddam envisages a way for politicised people on the first floor to 'exit the building', and pursue their causes through accepted mechanisms. Some theorists have also argued that 'polarisation' can have useful social outcomes (see e.g. Stewart *et al.*, 2020).

In fact, one might ask whether the 'unproblematic early stages' of the pathway (however defined) should feature at all in an account of the overall journey towards extremism. For our project, however, we have decided the early stages of the radicalisation process should be in scope. The main reason for this decision is the possibility that movement through the early stages of the pathway happens through *the same general mechanisms* that operate at later stages. If that is true, then studying the early stages of the radicalisation process can potentially illuminate what happens at later stages. Of course, that is a big if. But the social identity account of radicalisation certainly seems to posit the same basic mechanisms operating at each stage. In van Stekelenburg's account in particular, every stage in the radicalisation process is construed as involving 'tough identity work', to cognitively conceptualise relevant social groups, and to actively co-create them in the world, and to distinguish them from other groups. This work clearly takes different forms at different points in the pathway: but nonetheless, if there are generalisations to be made, we certainly want to make them (and

parsimony suggests we should make them).

In this report, we will describe mechanisms that operate in the same way at each stage of the radicalisation process as 'homogeneous'. Whether any of the mechanisms involved in radicalisation are homogenous is of course an empirical question. But accounts like van Stekelenburg's give some cause to think that there may be some degree of homogeneity in the mechanisms leading to radicalisation. The concept of homogeneity is particularly relevant to the main focus of our project, which is recommender systems. Recommender systems are certainly homogeneous in one trivial sense: the same recommender system operates on the user at each point in time, regardless of the user's behaviour. Of course the recommender system implements a learned function that maps a given user's platform history onto predictions about content preferences—and the user's platform history clearly changes over time. Whether the learned function is homogeneous is an empirical question, which we will return to in Section 5.7.2. But there are certainly reasons for thinking it might be. We will spell out these arguments in more detail in Sections 2.3.6 and 2.4. For now, our main point is that the *possibility* of homogeneous mechanisms in the radicalisation process is enough to keep the early stages of the process in scope for our study.

2.3.6 Possible roles for social media in the pathway towards extremism

The review just given emphasises the 'tough identity work' that people must do in order to become extremists, at all stages of the process. A key point to highlight is that social media platforms provide many ways of supporting this 'identity work'. In this section we examine social media platforms in the light of the above review, to ask what mechanisms they may provide to support the radicalisation process.

A social media platform is a powerful tool for the formation and maintenance of social groups (Backstrom *et al.*, 2006; Brady *et al.*, 2020). Crucially, social media platforms let users *actively participate* in groups, through a number of mechanisms designed specifically for the purpose: users can send posts to groups, or engage with other people's posts, both positively and negatively, through likes and dislikes, or through free-text comments; and users can resend comments to their own networks. In the account of radicalisation outlined above, active participation in groups plays a crucial role: people first identify themselves as belonging to a group (through a perceptual process), and then through various types of active participation, identify increasingly closely with that group, discriminate increasingly strongly against outgroups, and eventually participate in practices that isolate the group, and instil fear of external groups.

Of course, the vast majority of the time, these mechanisms help users create socially beneficial groups, and help people to find, join, and participate in these groups. Nonetheless, it may be that *alongside* these socially beneficial functions, social media systems also have small effects in encouraging users to move along the pathway towards radicalisation. That is the 'prima facie concern' we articulated in Section 1.6, and that motivates our project. In Section 2.4, we will look in more detail at what these 'small effects' might be.

2.3.7 Studies of the role of social media in radicalisation

In this section, we briefly review some studies of the role of social media in the radicalisation process. We begin with a group study, and then consider some more qualitative studies of individuals.

Frissen (2021) show that jihadist information seeking online was a significant direct predictor for 'cognitive radicalisation' in a cohort of Belgian young adults. The study focuses on jihadist extremism. Cross-sectional analysis of 1872 Belgian young adults shows that the most violent and radical materials were sought after but were the least predictive for radicalisation. Conversely, the static extremist magazines were sought by a small minority but were the strongly associated with radicalisation. The authors conclude that it is the consumers of extremist content who actively seek out such media, rather than the internet being the first to recommend.

Another way of exploring the effects of social media on radicalisation is to ask actual extremists whether social media played a role in their own radicalisation journey. An interesting study of this kind was performed by Baugut and Neumann (2020). These researchers used a qualitative research paradigm, conducting in-depth interviews with 44 'radicalized Muslim prisoners and former Islamists' in Germany and Austria. Respondents'

answers emphasised the role of structured online propaganda in online radicalisation. Respondents typically identified online platforms (particularly YouTube) as the place where they first had contact with radical content. Respondents also frequently recalled following links to recommended content. Often, the extremist intent of material from a given source only became clear after links were followed, which suggests deliberate structuring of radicalising content.[4]

2.4 Pathways towards harmful content: potential roles for general cognitive biases

Our account of the radicalisation pathway in Section 2.3 was primarily a psychological one, focusing on the cognitive processes of individuals moving towards radicalisation, and how these play out in the world. In this section, we will review a number of cognitive biases that have been claimed to operate in the process of radicalisation, and that are also implicated more generally in accounts of pathways towards 'harmful content'. These biases are operative in every aspect of people's lives: they certainly don't just operate in online or social media contexts. However, their manifestation in social media domains is of particular relevance, because in these domains they are registered by recommender systems, in ways that may systematically influence the subsequent behaviour of these systems, as we reviewed in Chapter 1. In this section, we will review five kinds of cognitive bias, to flesh out the concerns raised in Chapter 1. We will focus on evidence for these biases in social media domains, because their presence in social media is the key issue of concern. Concretely, we will review evidence that social media users *preferentially engage with*, and *preferentially disseminate*, content of certain kinds.

2.4.1 A bias towards 'moral emotional expressions' in political messages

Some interesting recent work has focused on biases in the transmission of political messages on social media. The key finding here is that political messages that contain 'moral emotional expressions' diffuse more readily on social media than other political messages. The original study showing this effect (Brady *et al.*, 2017) used Twitter's API. A corpus of 500,000 political messages on several current topics was gathered, and the frequency of moral, emotional and moral-emotional words in each tweet was computed, using existing lists of moral and emotional words. ('Moral' and 'emotional' words were those only appearing in the moral and emotional list respectively; 'moral-emotional' words appeared in both lists.) The number of retweets for each message was also retrieved, and a regression model was created to predict the retweet rate of a message from the number of words of each category it contained. There was no effect of moral words or emotional words on retweet rate, but a dramatic effect of moral-emotional words: adding a single moral-emotional word increased its retweet probability by 19%. This study has now been replicated many times; Brady and van Bavel (2021a) recently conducted a pre-registered replication on a larger set of tweets (800,000), along with a meta-analysis on 27 studies, and showed a consistent effect of moral-emotional words: on aggregate, they conclude that each additional moral-emotional word in a message increases its probability of being shared by 12%.

Brady *et al.* (2017) termed the effect of moral-emotional language on dissemination rate of messages 'moral contagion'. A related term, 'emotional contagion', was used by Kramer *et al.* (2014), to refer to their finding that exposure to emotional material in users' social media feeds influences the emotional content of their posts. Kramer *et al.*'s study involved a large-scale intervention on 600,000 Facebook users, that manipulated the recommender system that curated their news feeds to change the emotional content of items received, over the course of a single week. In one group of users, items with positive emotional content were omitted from the feed (stochastically, with a range of probabilities) during this time; in the other group, items with negative emotional content were omitted (by the same stochastic criterion). Regression models predicted the percentage of positive emotional words produced by users in the first group, and the percentage of negative emotional words produced by users in the second group during the treatment period. These models showed small but significant effects: users receiving less positive emotional content produced fewer positive emotional words (and more negative ones), while users receiving less negative emotional content showed the opposite

[4] The Royal Commission of Enquiry into the Christchurch attacks also reached various conclusions about the role of the Internet in the radicalisation of the attacker in its report—but since this relates to a single individual, we will not consider it here.

pattern. Even though the effects are small, they certainly compound the effects found by Brady and van Bavel.

2.4.2 A bias towards 'moral outrage' and negative emotions

Brady and van Bavel's (2021a) review singles out *negative* moral emotions as being particularly detrimental to relations between political groups. They cite another recent study (Brady and van Bavel, 2021b), showing that when users in a given political group express negative moral emotions, this consolidates their status as group members to other members of the group, and makes members of the opposing political group view them as 'less worthy of conversation'. Another recent study by Brady *et al.* (2021) shows a related compounding effect: users who receive positive feedback (from their in-group) for posts containing negative moral emotions subsequently produce more such posts. They suggest plausibly that basic reinforcement learning circuits might be responsible for this effect.

Negative moral emotions are the focus of an earlier article by Crockett (2017), who referred to them by the term 'moral outrage'. In this paper, Crockett cites a study showing that people are more likely to 'learn about immoral acts' from online sources than from traditional media or immediate experience, and moreover that immoral acts encountered online elicit stronger moral outrage than those encountered in person or via traditional media. These results suggest that online environments increase the virality of negative moral emotions. Facebook's ex-employee Frances Haugen referred to internal studies reaching similar conclusions: in her words, 'hateful, polarising content gets more distribution, more reach'.

A related finding is that participants in online discussions have a tendency towards making negative comments that increases with their level of participation in the group (del Vicario *et al.*, 2016). The same study found that the more active participants moved faster towards negative sentiments than less active participants. This study examined discussion in a 'conspiracy theory' domain as well as one in a scientific domain, and found the same effects in both domains. The finding about the scientific domain shows that 'emotional contagion' doesn't just happen in domains where morality is at issue.

The virality of negative moral emotions is likely to be dependent on many factors that are only just beginning to be explored. There may be cross-cultural variation, or variation between political groups. Yoshida *et al.*'s (2021) study of Japanese Twitter bears on both these issues. These researchers found that conservative messages on Twitter are more effectively communicated to moderate users than liberal messages. They also showed that while there were no differences in the frequency of hashtag use in conservative and liberal tweets, 'emotion words conveying dislike' were more frequently used in conservative tweets. It may be that this difference is what causes conservative tweets to be more efficiently communicated, though more research would be needed to test this idea explicitly.

2.4.3 A bias towards content about a political 'out-group'

Another bias that has been studied in the political sphere is a bias in favour of posts which are *about* a political out-group. Rathje *et al.* (2021) report that posts about political opponents are substantially more likely to be shared on social media. They also note that this outgroup effect is much stronger than other established predictors of social media sharing, such as emotional language.

Rathje *et al.* also found that outgroup language elicited particular effects on recipients. Language about the out-group was a very strong predictor of 'angry' reactions (the most popular reactions across all datasets), and language about the in-group was a strong predictor of 'love' reactions, reflecting in-group favoritism and out-group derogation. This out-group effect was not moderated by political orientation or social media platform, but stronger effects were found among political leaders than among news media accounts. This effect was shown on both Facebook and Twitter.

All these effects are likely to contribute to the mechanisms that lead to political polarisation. As discussed in Section 2.3.3, these mechanisms involve intertwined processes of increasing identification with an in-group, and increasing hostility towards an out-group, both in the sphere of psychological conceptualisation and in the sphere of collective public actions.

2.4.4 A bias towards falsehoods

Another bias that has been identified online is for the propagation of false information, rather than true information. The key study here is by Vosoughi *et al.* (2018). These researchers investigated the differential diffusion of all the verified, true and false news stories on Twitter from 2006 to 2017. The data comprised approximately 126,000 cascades of news stories spreading on Twitter, tweeted by about 3 million people over 4.5 million times. The news pieces were categorised as true or false using information from six independent fact-checking organisations that exhibited 95%–98% agreement on the classifications. It was found that falsehoods diffused significantly further, faster, deeper, and more broadly than the truth in all categories.

The effects were most pronounced for false political news than for news about terrorism, natural disasters, science, urban legends, or financial information. Robots accelerated the spread of true and false news at the same rate, implying that humans, not robots, are more likely responsible for the dramatic spread of false news.

Connected with this bias towards falsehoods, there also appears to be a bias towards *surprising information*. False information is sometimes also surprising—and the elaborately bizarre false tenets of conspiracy theories are often particularly surprising. As reviewed in Loewenstein (2019) and Simandan (2020), people attend more to surprising facts, they are more aroused by surprising facts, they are better at remembering surprising facts, and they also find surprising facts more appealing. All these factors make them more likely to disseminate surprising facts online—including, presumably, facts that are surprising because they are false. (Of course, facts that contradict strongly held beliefs will elicit emotions other than surprise; surprises are associated with new discoveries.)

2.4.5 A bias towards 'sensational content' and 'clickbait'?

Another possible user bias that we mention more tentatively is one towards 'sensational' content. Sensational content can include emotional content, but it is broader than the category of 'moral' emotional content discussed above. We follow Mourao and Robertson (2019) in using Kilgo *et al*'s (2018) definition of sensationalist stories, as 'stories that intentionally [evoke] emotion in the beginning of the article, [use] extreme circumstances to grab attention, [simplify] and [trivialize] a complex topic, [promote] shock value, or [are] presented in a tabloidlike way'.

The category of sensational content thus defined overlaps with another relevant category, specific to online media, called 'clickbait'. Clickbait is a term that relates specifically to hyperlinks appearing in online media (or to the text snippets that accompany these), if they are designed to create 'curiosity gaps' in the minds of readers, that entice them to follow the links (Blom and Hansen, 2015; Kilgo and Sinta, 2016). By these definitions, 'clickbait' items are sometimes, but not always, 'sensational'—but they are close enough that we will treat them together.

Do users have a tendency to engage with 'sensational content' or 'clickbait'? The evidence in each case is mixed. Some studies, examining some media, find that users engage more with sensational content (see e.g. Tenenboim and Cohen, 2015). Some studies, examining some media, find that users engage more with clickbait (see e.g. Rony *et al.*, 2017). But Mourao and Robertson (2019), in a careful study of Facebook and Twitter, found no evidence that users engage more with sensational content or with clickbait. Having said that, there are persistent stories that sensationalism is a successful strategy for social media content providers. A recent example is a Wall Street Journal story about an internal study by Facebook researchers about its recommender algorithm (Hagey and Horwitz, 2021). According to this story, 'publishers and political parties were reorienting their posts toward outrage and sensationalism. That tactic produced high levels of comments and reactions that translated into success on Facebook'.

2.4.6 Possible effects of cognitive biases in social media

To recap: there is strong evidence that social media users are biased towards 'moral emotional expressions', leading to 'moral contagion' and 'emotional contagion' (Section 2.4.1), towards 'moral outrage' and negative

emotions generally (Section 2.4.2), towards content directed at political outgroups (Section 2.4.3) and towards false information (Section 2.4.4); and there is some evidence that users are also biased towards sensational information (Section 2.4.5).

As we have already noted, these biases operate quite generally in people's lives. In some cases there is evidence they are stronger online than offline (especially for 'moral outrage'). But our main concern is not with the strength of biases, as they are manifested online. Our main concern is about possible *consequences* of these biases, as they play out in social media contexts.

To spell out the concern concretely: the worry is that the biases shown by a user on some social media platform will be registered by the *recommender system* that learns about the user on that platform, and then chooses further content for the user in the light of this learning. In the scenario we are concerned about, the content it presents next for the user will include more moral emotional expressions, more moral outrage, more negative expressions, more out-group content, more falsehoods, and (perhaps) more sensationalist content. If the user *maintains* her biases in relation to this new material, the items she selects from within this new set will be *further* biased in the same direction. And, in turn, the recommender system will further revise its learning to accommodate these additional biases. In this scenario, the proportion of content shown to the user that includes moral emotions, moral outrage, out-group content, falsehoods and sensationalism will continue to grow over time.

The basic mechanism illustrated here is the positive feedback loop articulated in Chapter 1, through which users end up viewing an arbitrarily narrow range of content. But the presence of problematic biases in user preferences adds a new dimension of concern: the narrow space towards which users are progressively steered may be one dominated by content containing moral emotions, moral outrage, out-group references, falsehoods and sensationalism. The proliferation of such content is harmful in itself—but a particular concern is that it might have a role in encouraging users along pathways towards extremist positions. As summarised in Section 2.3, the pathway to extremism involves progressively heightened emotions, progressively polarised conceptions of in- and out-groups, and progressive adoption of countercultural narratives. There is a real concern that recommender systems, in the presence of the biases reviewed in the present section, may have a role in encouraging users along this pathway.

We should reiterate that the biases we are concerned about are quite small: naturally, a host of other factors, many positive, many idiosyncratic, will influence users' trajectories through social media content spaces. But small biases *that systematically perturb a dynamical system* may still have large effects on its outcome, and are still a matter for concern.

2.5 Interim summary

In this chapter, we began in Sections 2.1 and 2.2 by reviewing methodologies for identifying harmful content, and outlining a basic typology of 'harmful content' on the Internet. In Section 2.3 we gave a basic account of the pathway towards violent extremism and its associated harmful content, TVEC. Then in Section 2.4, we enumerated some cognitive biases that may push social media users in the direction of harmful content of several types, including possibly TVEC.

Sections 2.3 and 2.4 were intended to flesh out the 'prima facie cause for concern' with social media recommender systems, first noted in Section 1.6, by discussing possible roles for recommender systems in theoretical models of radicalisation processes. But this chapter is also intended to provide context for the upcoming Chapters 3 and 5), which review empirical methods for studying the effects of recommender systems on users' relationships with harmful content of various kinds. While our prima facie argument turned on formal models of recommender systems, the empirical methods we review in the next chapters focus on real systems, and real users. In order to study these, we must work with actual online content, and use methods that classify this content into the kinds of category surveyed in Section 2.2. To conclude the current chapter, and as a bridge to Chapters 3 and 5, we will review techniques for automatically classifying content generated by social media users.

2.6 Classification of harmful online content

Defining categories of harmful online content is one thing; *applying* these definitions to actual online content is another. A key issue is the need to classify content at scale. Whether harmful content is being identified by a hosting platform, as part of its moderation process, or by researchers investigating hypotheses about harmful content, it is typically infeasible to perform classification entirely by hand; some aspects of the classification process need to be automated.

In the last few years, automatic identification of harmful content has become a new field of computational linguistics (for textual content), and a new field of computer vision (for images and video content). In Section 2.6.1, we will briefly review the common methods currently used for classifying harmful content, and that will feature in the experimental methods we survey in Chapters 3 and 5. Methods for identifying false information are rather different, so we will consider these separately in Section 2.6.2.

2.6.1 Automatic classification of hate-related content

2.6.1.1 Initiatives in the academic community

There are several workshops devoted to the task of automatic classification of harmful content. The Workshop on Online Abuse and Harms (WOAH),[5] previously the Workshop on Abusive Language Online,[6] is collocated with the Association for Computational Linguistics conference, and focuses on text classification. The International Workshop on Cyber Social Threats (CySoc)[7] is collocated with the AAAI Symposium on Web and Social Media, and also considers classification of images and videos.

These conferences are a focus for the development and use of public resources relating to content classification. Language resources include lexicons of abusive expressions (e.g. Wiegand *et al.*, 2018), corpora hand-annotated for harmful content of various kinds (e.g. Zampieri *et al.*, 2019; Basile *et al.*, 2019), and publicly available classifiers (Aluru *et al.*, 2020); a good summary is given by Kiritchenko *et al.* (2021). In the usual AI style, hand-annotated corpora provide benchmarks for assessing the performance of classifiers, often in the context of 'shared tasks': 'HatEval' and 'OffensEval' are two recent examples (see again Zampieri *et al.*, 2019; Basile *et al.*, 2019).

A good illustration of current resources for hate speech detection is the HATECHECK suite of tests (Röttger *et al.*, 2020). The suite is a set of functional tests for hate speech detection models. The tests were chosen through interviews with civil society stakeholders and a review of previous hate speech research. The functional differences between hateful and non-hateful content can be challenging to detection models. For each functional test, sets of target test cases were validated through a structured annotation process. HATECHECK was demonstrated as a diagnostic tool which revealed weaknesses of recent transformer models as well as two commercial models for hate speech detection. Models were shown to be overly sensitive to particular keywords and phrases, as evidenced by poor performance on tests for reclaimed slurs, counter speech and negated hate. The transformer models also exhibited strong biases in target coverage.

2.6.1.2 Initiatives by tech companies

Tech companies use their own in-house resources to detect harmful content. The classifiers and training sets are not typically made publicly available—we presume this is in part due to IP issues, and in part because full disclosure of these resources would allow the development of adversarial methods for avoiding content detection. However, some companies have placed some resources relating to harmful content classification in the public domain.

[5] https://www.workshopononlineabuse.com/
[6] https://www.aclweb.org/portal/content/3rd-workshop-abusive-language-online
[7] http://cysoc.aiisc.ai/

Google's 'Perspective' API Google's public 'Perspective' API allows external users to access a classifier that classes short texts into various categories of harm-related content (including 'threat', 'sexually explicit', 'insult' and 'severe toxicity').

Facebook's classification technology Facebook are somewhat more open about classification technology—not about their classifiers *per se*, but about systems that support these classifiers. Their Reinforcement Integrity Optimiser (RIO)[8], a system that has been used since to optimise hate speech classifiers vetting Facebook and Instagram uploads. Their SimSearchNet[9] convolutional neural network is used to help detect near-identical duplicated content, to ensure content classification labels are applied as widely as they should be. (A key use case is in labelling harmful misinformation regarding COVID-19.) Their Whole Post Integrity Embeddings (WPIE)[10] system analyses platform content in a multimodal manner, for example considering text and image content together, rather than separately, which was the standard approach until recently. Finally, their XLM-R model[11] (Conneau *et al.*, 2020) facilitates cross-lingual content classification, with the goal of using training data in one language to classify texts in another language. Hate speech is one target domain for this technology—though we note that many types of hate speech differ across cultures, and hence across languages.

We will discuss what we know about the performance of companies' current harmful content classifiers in Annex B. The effectiveness of these classifiers is another closely guarded secret. But we are fairly sure current methods have relatively low performance. (For instance, one of the Facebook documents recently leaked by Frances Haugen reports an internal study estimating Facebook 'may action as little as 3–5% of hate and about 6-tenths of 1% of V & I [violence and incitement] on Facebook despite being the best in the world at it'.)

Of course, for high-profile social media users (and especially for politicians), moderation of harmful content is done by hand, through processes that resemble the decisions made by human newspaper editors (see e.g. Conger and Isaac, 2021). Rather different standards are applied in these cases; for Twitter, for instance, the 'newsworthiness' of a Tweet and whether it 'in the public interest' determine whether the normal rules will apply.[12]

2.6.1.3 Classification methods used in studies of social media

In studies of harmful content on social media, classification is normally done using methods that are somewhat simpler than full content classifiers. Studies of textual content often use extremely simple methods: for instance, as discussed in Section 2.4, we mentioned several studies that analyse texts by the presence of particular words or expressions, without regard for context or syntactic structure (Brady *et al.*'s 2017 study of 'moral' and 'emotional' utterances is a good example.) Studies also use analyses based on hashtags, or based on the identity of content providers. Inaccuracies in these methods can often plausibly be regarded as noise. Classification errors in *observational studies* of social media systems are unlikely to be as harmful as they are when used to actively moderate content, for instance.

2.6.2 Automatic identification of false information

2.6.2.1 Initiatives in the academic community

Automated methods for assessing the truth of online content are the focus of a different part of the AI community, concerned broadly speaking with 'information extraction'. A key conference here is the FEVER

[8] https://ai.facebook.com/blog/training-ai-to-detect-hate-speech-in-the-real-world/

[9] https://ai.facebook.com/blog/using-ai-to-detect-covid-19-misinformation-and-exploitative-content/

[10] https://about.fb.com/news/2019/11/community-standards-enforcement-report-nov-2019/

[11] https://ai.facebook.com/blog/-xlm-r-state-of-the-art-cross-lingual-understanding-through-self-supervision/

[12] https://twitter.com/Policy/status/912438226736041985

(Fact Extraction and VERification) conference.[13] Again, work in the academic community centres around the curation of public datasets, and participation in 'shared tasks' that reference these datasets. A good survey of current work is given by Oshikawa *et al.* (2020).

2.6.2.2 Initiatives by tech companies

Again, tech companies keep their methods for identifying 'fake news' mostly to themselves. But we believe they maintain their own databases of trusted news sources (see e.g. Mosseri, 2018), and we know they work alongside partner news providers in some contexts (see e.g. Timmins, 2021). We also know that they contract a great deal of fact-checking work out to external organisations. A good recent review of fact-checking mechanisms used by social media platforms is given in the Broadband Commission's recent report on disinformation (Bontcheva and Posetti, 2020).

[13] https://fever.ai/FEVER

3 A review of 'external' methods for studying the effects of recommender systems on users

In this chapter, we will survey the findings of studies examining the effects of social media recommender systems on the users who consume their recommendations. We focus on studies conducted 'externally' to the companies that deploy recommender systems. These studies use data that are available outside social media companies, either through APIs provided by companies, or through direct experiments on platform users or user interfaces. We also focus on studies that explore whether the operation of social media platforms has any harmful effects on users or user communities. The harmful effects in focus are those reviewed in Chapter 2: they include measures taken over communities (relating to polarisation), and measures taken over individual users (relating to their access of harmful content of various kinds).

Our review supplements recent comprehensive reviews by UNESCO (Alava *et al.*, 2017) and by the GIFCT (GIFCT, 2021c). The main difference in our review is that it is narrower than the UNESCO review, focussing mainly on quantitative methodologies, and that it organises studies by the different methodologies available to external researchers.[1] We consider studies of differences between whole populations (Section 3.1), studies using logging software on user machines (Section 3.2), studies using company-supplied APIs and datasets (Section 3.3), studies using 'robot users' of social media systems (Section 3.4), and intervention studies on social media users (Section 3.5).

3.1 Population-level studies

Population studies seek to examine trends in social attitudes in relation to trends in social media (or Internet) usage, and also in relation to broader social and political trends. Methodologically, a key focus of these studies is to look for 'off-platform' factors that may account for trends in social attitudes. We will base our review on two papers by Levi Boxell and colleagues, that focus on trends in polarisation, and on a review of these papers by Steinhardt (2021).

Boxell *et al.* (2020) present a very useful broad study of 'affective polarisation' across 12 OECD countries, and over a long period of time (the past 40 years). They define affective polarisation as 'the extent to which citizens feel more negatively toward other political parties than toward their own'—a definition they attribute to Iyengar *et al.* (2019). The focus for Boxell *et al.*'s study is on possible explanations for changes in polarisation. They consider four possible categories of explanatory factor, relating to the economy, to technology, to demographics and to politics. A key finding in Boxell *et al.*'s analysis relates to the 'technology' factor. Within a focus group of nine Western countries, they find fairly uniform trends in Internet and broadband penetration over the last 40 years. But the trends in affective polarisation are not uniform across countries: in particular, polarisation has only increased since 2000 in the US, Canada and the UK, while for other countries is is relatively flat by their measures. This is prima facie evidence that factors other than the Internet are driving polarisation. But it is worth noting some potential confounding variables here: in particular, the US, Canada and the UK are the only countries in the study with a two-party political system, and measures of polarisation in other countries may not be comparable.

Another study by Boxell *et al.* (2017) makes demographic comparisons rather than regional ones. This study focuses on the US, and compares US populations of different ages on a number of factors, including Internet use and political polarisation. Boxell *et al.* find that older people use social media (and the Internet in general) less than younger people. They also find that political polarisation has increased more rapidly among older groups than among younger groups. Again, this result seems to argue against social media as a primary driver

[1] We will not focus on studies that rely on 'self-report' of user effects, which have their own methodological problems (see Griffioen *et al.*, 2020 for a good discussion).

of political polarisation in these populations. But again, the argument isn't clearcut: there could be other variables that distinguish these groups that independently impact on polarisation. In particular, there is good evidence that older people are more inclined to believe what they see online (see e.g. Brashier and Schacter, 2020 for an especially relevant review). Confounding variables are often an issue for population studies, as we will discuss more in Section 3.6.

Boxell *et al.*'s (2020) study also aims to find non-technological factors that positively correlate with changes in affective polarisation, that may provide a better explanation than technical factors. They find few significant correlations, but there are some: for instance, the proportion of 'non-white' people in a population is significantly correlated with its polarisation. Of course, no-one would expect that Internet or social media use is the *only* factor influencing polarisation. Social media may not even be a *large* influence, compared to other factors. (In the US, for instance, trends in TV news coverage are likely a larger influence; see e.g. Prior, 2007; Hmielowski *et al.*, 2016.) Our focus on possible effects of social media is purely because such effects *may be due to AI mechanisms*, which puts them in scope for our GPAI working group.

Even though Boxell *et al.*'s (2020) study does not find strong effects of any one variable, it nonetheless makes a very important point, that we will return to several times in our report: trends towards polarisation, and likely in other problematic directions, are heavily dependent on time, and on place, and often also on particular populations. This basic result is likely to hold for studies specific to social media platforms as well: for instance, we may see a worrying trend in *some* platforms, in *some* countries, at *some* times, but not at others.

3.2 Studies using logging software

Logging software can be installed on volunteer users' Internet browsers, to study their browsing habits. The data gathered by browser loggers inform about how users behave on social media platforms—but they also inform about users' behaviour on other Internet sites. While the population-level methods discussed in Section 3.1 measure people's behaviour in a very broad spheres, going well beyond Internet use, browser logging studies measure a narrower domain of 'off-platform' behaviour, that is away from social media sites, but still related to Internet use. Logging studies are also much finer grained, in that data about *individuals* is preserved in logs. Flaxman *et al.* (2016) studied consumption of online news content by analysing web browsing histories for users of Microsoft's Internet Explorer. Histories were collected through Explorer's 'Bing' Toolbar, which users can optionally install: when they do so, they can volunteer to share their browsing data for research purposes. A key finding for Flaxman *et al.* was that the vast majority of users' consumption of news content came from users visiting their favourite (typically mainstream) news outlet. It was relatively rare for users to access news content by clicking on links offered on social media sites. Nonetheless, populations who accessed news through social media were found to be more 'politically segregated' in the sources of news they consulted than populations who consumed news directly from news sites—especially for access of 'opinion pieces' rather than descriptive news. This effect is not large, but it is nonetheless significant—and we should bear in mind that accessing news through social media is more common today than it was in 2016 (see e.g. Shearer and Mitchell, 2021).[2]

A key concern for any dataset derived from logs of volunteer users' online behaviour is whether the set of users is fairly sampled. Not all users want to have their online activity monitored, even under guarantees of privacy and anonymity. However, some sampling methods are likely to be more representative than others. A logging dataset from the Nielsen market analytics firm is particularly useful, as all sampled users were paid for their participation. Hosseinmardi *et al.* (2020) used this dataset for their study of YouTube. These researchers showed that 'news-related' content on YouTube accounted for only 11% of content accessed by users; again, mainstream media sources accounted for the majority of news-related content. They did find that the proportion of users 'strongly engaged' with far-right YouTube channels rose from 2016 to 2020; however only a small proportion of far-right video views can be plausibly attributed to YouTube's recommender

[2] Flaxman *et al.*'s study also found that accessing news through social media increased individuals' exposure to material from the opposing political camp. They found this result 'counter-intuitive'—but as we have already discussed, polarisation involves forming attitudes towards the 'out-group' (see Section 2.3.3) and social media use encourages references to this out-group (see Section 2.4.3).

algorithm. By this the authors mean that consumers of far-right content arrive via multiple 'pathways' such as search and external websites, as well as previously watched videos. In particular, they do not see any trend *within single YouTube sessions* in the direction of far-right content. By this the authors mean that the rate of consumption of far-right content does not increase either over the course of a session or with the length of a session, as would be expected if users were being actively steered to such content via their in-session recommendations.

Hosseinmardi *et al.*'s study provides quite strong evidence that YouTube's recommender system does not have a large role in encouraging users to move towards far-right content: they attribute the rise in users' interest to 'off-platform' factors. But we should reiterate that even small effects of a recommender algorithm are of interest—especially if they can be eliminated by modifying the algorithm. We also note that Hosseinmardi *et al.*'s way of measuring the role of recommender systems involves some inference beyond the data: as they also concede, their logging software does not show the content actually recommended to users, and so doesn't register user clicks on recommended items. This data is most readily obtained in 'platform-internal' studies, of the kind we will discuss in Chapter 4.

3.3 Studies using public APIs and datasets, and private analytics

Social media platforms have always made data about their operation available externally, through APIs. In the wake of the Cambridge Analytica scandal, companies reduced what information is accessible through APIs (see e.g. Tromble, 2021). There are also often limits on the number of requests that can be made. Both limitations can be reduced by commercial users who pay for access.

Social media companies have also created many datasets for use by academic researchers interested in studying their operation (and studying society more broadly). For instance, the Cambridge Analytica scandal also led Facebook to establish Social Science One, a partnership between Facebook and public organisations, whose aim is to find ways for Facebook to share its data with external groups.

There have not been many studies using APIs to explore performance of recommender systems directly, because APIs do not surface much detail about recommender system behaviour. We will describe the best known direct study of recommender systems in Section 3.3.1. Most API studies reveal broader trends in platform use, charting data similar to the population studies described in Section 3.1, but with finer-grained detail. One of these studies directly bears on hypotheses about recommender systems; we will describe it in Section 3.3.2. In Section 3.3.3, we describe an example study using commercial access to APIs, and in Section 3.3.4 we describe the current state of play with Social Science One. We conclude in Section 3.3.5 with a study of recommender systems that was conducted internally to a social media company (Facebook), rather than using an external API. (We include this study in the current chapter because it doesn't exploit the internal methods that are our focus in Chapter 4.)

3.3.1 Ledwich and Zaitsev's study of recommendations using YouTube's API

YouTube's API provides number of subscribers and aggregate views for each YouTube channel. It also indicates for each channel what other channels are recommended. Ledwich and Zaitsev (2019) present a study that classifies 800 political YouTube channels into classes on the left-right spectrum, and then charts the nature of recommended channels. Their key finding is that recommendations do not tend to move in the direction of radicalised content. On the contrary, the authors found that YouTube's recommendation algorithm actively discourages viewers from visiting radicalising or extremist content. The algorithm favours mainstream media over independent channels and there is favour towards left-leaning or politically neutral channels.

However, Ledwich and Zaitsev's study is significantly hampered by the limitations of YouTube's API. As they note, the API only provides recommendations for an 'anonymous' account, that has no videos in its history. (They have to supplement API data with data gathered from their own scraper, which had a similarly blank history.) They concede that the recommendations offered to a specific user, with a specific history, may differ from those offered to a generic account. But they argue that this is unlikely, based on a description of the

recommender algorithm offered by Zhao *et al.* (2019). As they say, '[i]t would seem counter-intuitive for YouTube to apply vastly different criteria for anonymous users and users who are logged into their accounts, especially considering how complex creating such a recommendation algorithm is in the first place'. But this argument is flawed. The whole point of a recommender algorithm is to recommend different items for different users, based on an extended history of engagement with the platform—and indeed, the algorithm described in Zhao *et al.* (2019) clearly takes many user-specific features of this kind into account. As also noted by Whittaker *et al.* (2021), the whole filter bubble hypothesis is predicated on the idea that recommender algorithms learn about *individual users*. We question whether strong conclusions about recommender systems can be drawn from a study that doesn't incorporate this assumption.

3.3.2 Munger and Philips' study

Munger and Philips (2019) use YouTube's API to study population trends on the platform, rather than recommendations. But their work is still very relevant to work on recommender systems, because their aim is to demonstrate patterns of behaviour that are unrelated to recommender systems. In particular, they focus on YouTube's status as a 'media company', that pays creators of video content, and that supplies a very large choice of 'channels'. The former property is attractive to 'suppliers' of content, and the latter is attractive to 'consumers' of content; they suggest that principles of 'supply and demand' on YouTube may outweigh effects of its recommender algorithm. They argue that rises in the amount of extreme content on YouTube may be better explained by supply and demand principles. On their hypothesis, a demand already existed for extreme content, that wasn't catered for by mainstream media outlets, but was readily met by YouTube providers who were remunerated for their content. In support of this analysis, they use API data to show that viewership of far-right videos peaked in 2017, and has since been in decline, while the increases in viewership since 2017 have mainly been due to the arrival of 'mainstream conservatives' as content providers. They argue these findings are better explained on a supply-demand account than as an effect of a recommender system.

Munger and Philips' supply and demand analysis of dynamics on YouTube is a useful complement to analyses relating to recommender systems. Their data provides interesting, if indirect, evidence for supply and demand factors. But we don't see it as providing strong evidence *against* a role for recommender systems. One might expect effects due to recommender systems to be constant over time, in some sense—but companies frequently modify their recommender systems, in ways that are not made public, and frequently change their content moderation policies too. Recommender systems could still be responsible for some component of the variations charted by Munger and Philips.

3.3.3 Allcott *et al.*'s study

As noted above, social media companies provide free APIs surfacing some information about the operation of their platforms—but more detailed information is sometimes available at a price. Some information can be bought directly from companies: for instance, Twitter allow clients to buy detailed data about the number of times tweets are shared. Other information can be acquired from data brokers of various kinds. Allcott *et al.* (2019) use data from BuzzSumo, a commercial database that tracks user engagement with Internet content across a variety of platforms, using a mixture of free and paid-for data. Using this data, they surface some interesting population-level statistics about the diffusion of misinformation on social media, which is their topic of interest. In particular, they show that user interactions with false content rose on both Facebook and Twitter from the start of 2015 to the end of 2016. Since then, however, interactions with false content fell relatively sharply on Facebook, while they continued to rise on Twitter until the end of the study period (July 2018). This again demonstrates that different platforms exhibit different trends, as already noted in Section 3.1. Again, changes to platform algorithms are likely responsible for at least some of this variation.

3.3.4 Studies using Social Science One

Social Science One, hosted by Harvard's Institute for Quantitative Social Science, was established with the vision of becoming a hub of academia-industry partnerships for privacy protected information sharing for social good. The first collaboration was with Facebook with the aim of compiling a large public dataset of Facebook

URLs for scholarly analysis. Due to slow initial progress, the Co-Chairs and European Advisory Committee of Social Science One made a public statement in 2019 in which they declared that they were considering stepping back from the initiative (de Vreese *et al.*, 2019). However, this open declaration seems to have expedited the process, and the first *URLs Light Dataset* was released in early 2020, closely followed by the *URLs Full Dataset*.

The dataset is the largest of its type and contains a total of more than 10 trillion numbers that summarise information about 38 million URLs shared worldwide more than 100 times publicly on Facebook (between 1/1/2017 and 7/31/2019). It also includes characteristics of the URLs (such as in which country they were shared and whether they were fact-checked or flagged by users as hate speech) and the aggregated data concerning the types of people who viewed, shared, liked, reacted to, shared without viewing, and otherwise interacted with these links. The dataset is protected by the principles of 'differential privacy' by adding specially calibrated noise (see Evans and King, 2020). The noise guarantees that individuals who may be represented in the data cannot be re-identified, and any clicks, shares, or other actions cannot be associated with any one person. Despite the noise, differential privacy makes it possible for statistical analysts to learn social science patterns from the same data. However, as far as we know, Social Science One does not yet provide information about the sets of recommendations presented to users, or the items users selected within these sets. So the dataset can't yet be used directly for studying recommender system effects on users. Even if recommendation information were present in the dataset, the noise added to it might make it hard to track individual users' interactions with recommender algorithms.

3.3.5 Bakshy *et al.*'s study of information flow on Facebook

Bakshy *et al.*'s (2015) study is interesting, because it was done in collaboration with Facebook, using data not available outside the platform. Facebook, of course, is concerned to explore possible harmful effects of its recommender systems on users, as much as external stakeholders are—and the same is true of other social media companies. We place Bakshy *et al.*'s study in the current 'external methods' chapter, because it doesn't exploit the methods companies normally use to examine recommender system effects; its methods have more in common with the API methods discussed in the current section.

Bakshy *et al.*'s study explored factors leading to political polarisation on Facebook. Their investigation focused on US users' exposure to news items appearing on other websites. A population of users was gathered who disclosed their ideological affiliation, on a scale from -1 (liberal) to +1 (conservative). A set of 'hard' news stories concerning politics and world affairs was also gathered. Each story was evaluated on the same liberal-conservative scale, by averaging the affiliation of users who had 'shared' the story with friends.

Bakshy *et al.* examined two factors influencing how a Facebook user receives links to stories. One is the user's group of friends: items from the user's friends predominate in her news feed. The other is Facebook's ranking algorithm, that selects which 'potential' feed items to 'expose' to the user, and what order to present these in. The first factor was found to be highly important. As a general effect, people tended to be friends with others who shared their political perspective (the 'homophily' effect we noted in Section 1.3.2), and liberals tended to share liberal stories, and conservatives tended to share conservative stories. The political complexion of stories received by a given user thus depends in a fundamental way on her friends. Bakshy *et al.* nonetheless found a certain amount of 'cross-cutting content': liberal stories shared by conservatives and vice versa). The second factor was found to be far less important. Bakshy *et al.* computed the amount of cross-cutting content in (i) shares from a random sample of users, (ii) shares users were sent by their actual friends, which were 'potential' feed items, and (iii) shares users actually saw, because they were selected by the recommender system. There was a large difference between measures (i) and (ii), but only a small difference between measures (ii) and (iii), as shown in Figure 3.1.

Bakshy *et al.*'s study makes it clear that users' choice of friends is far more important than the feed ranking algorithm in determining the complexion of political material they are exposed to. But the recommender algorithm is also shown to play a role, albeit a smaller one. Recall that small effects of recommender systems are still of considerable interest, because they are perturbations of a dynamical system: over time, they can add up. Recall also that the process of choosing friends is an integral part of the pathway towards extremism: as discussed in Section 2.3, the formation of social groups is interleaved with the process of forming opinions. It

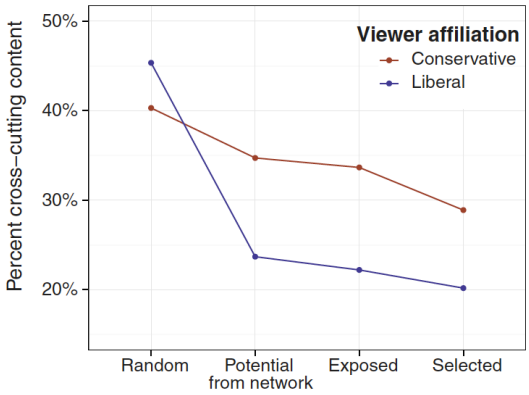

Figure 3.1: Factors influencing the amount of cross-cutting political content in Bakshy *et al.*'s (2015) study, showing the effect of friend group ('potential'-'random') compared to the effect of the recommender system ('random'-'exposed'). Results are broken down by affiliation of users (liberal or conservative).

is also worth bearing in mind that Facebook makes recommendations to users about potential friends ('people you may know'), as well as about content items. The way these recommendations are made is not known, and certainly wasn't part of Bakshy *et al.*'s study.

3.4 Studies of automated followers of recommender system suggestions

As already noted, using public APIs to study platform recommendations makes it hard to study how recommendations are *personalised* to individual users. (For instance, Whittaker *et al.* (2021) note that APIs don't allow study of individual users' 'relationship' with a given platform.) To circumvent this problem, another type of study builds automated systems that engage with social media accounts in the place of human users: what we might call 'robot' social media users. These automated users are assigned to follow recommended links, in one way or another: at issue is where such robotic users end up.

Often in these studies there is an initial phase where a set of channels for study is identified manually by researchers, following recommender system links, along with links recommended by content authors. After this, automated users can be assigned to channels, implementing algorithms that follow links, and gather a large dataset of content and recommendations for analysis. A good example is Ribeiro *et al.*'s (2020) study of YouTube. The study analyzed 330,925 videos posted on 349 channels, classified into four types: Media, the Alt-lite, the Intellectual Dark Web (I.D.W.), and the Alt-right. The authors show that the three non-media channel types majorly share the same user base and that these users gradually move from mild to more extreme content. The authors find that Alt-lite content is easily reachable from I.D.W. channels, while Alt-right videos are reachable only through channel recommendations. These authors conclude that YouTube's algorithm does have a role in migrating users towards more extreme content.

Another example of robot users is Papadamou *et al.*'s (2021) study of random walkers in the domain of 'incel' content on Youtube. As noted in Section 2.2.3, these authors find that during the last decade, the volume of incel-related videos and comments has increased considerably. More alarmingly, Papadamou *et al.* also suggest that the platform may also play an active role in steering users towards such extreme content. The authors found that users have a 6.3% chance of being suggested an incel-related video by YouTube's recommendation algorithm within five hops when starting from a non incel-related video. Also, users who have seen two or three incel-related videos at the start of their session see recommendations that consist of 9.4% and 11.4% incel-related videos, respectively. Furthermore, the portion of incel-related recommendations increases substantially as the user watches an increasing number of consecutive incel-related videos.

Automated users were also deployed in Cho *et al.* (2020)'s study of YouTube, where the automated agents were created to approximate the behaviour of volunteer students, based on their responses to a questionnaire. The experiment examined whether exposure to algorithmically recommended content reinforces and polarises political opinions. Results suggested that political self-reinforcement, as indicated by the political emotion-ideology alignment, and affective polarisation are heightened by political videos selected by the YouTube recommender algorithm, based on participants' own search preferences. The authors concluded that the YouTube algorithm does have an effect but it is heavily influenced by the searcher's preconceived preferences: in their judgement, the algorithm does not cause the inception of this radicalisation.

A final useful example of automated users comes from Whittaker *et al*'s (2021) study of YouTube, Reddit and Gab. Whittaker *et al.* started three accounts on each of these platforms, each subscribing to a mixture of far-right and apolitical channels. An 'extremist interaction account' modelled a user predisposed to pick mainly far-right recommendations, and a 'neutral interaction account' modelled a user predisposed to pick mainly neutral channels. Both accounts engaged automatically with each platform twice a day for a two-week period; a third 'baseline account' modelled an inactive user during this same period. Whittaker *et al.* examined how much far-right content was recommended to each account at the end of this period. They found that YouTube recommended significantly more far-right content to the 'extremist' account than to the other two. Interestingly, this effect was not observed either on Reddit or on Gab (for which a slightly smaller dataset was gathered). The difference between YouTube and Reddit is striking: it suggests that differences in the design of recommender systems across platforms can have significant influence on users.

3.5 Studies that intervene in social media users' actual behaviour

A final way of assessing the effects of social media systems is to conduct controlled studies asking volunteer users to behave in particular ways in relation to social media.

The most dramatic of these studies ask a group of users to abstain from social media altogether, and compare the attitudes of these users to a control group who use it as normal. A well-known study of this kind was conducted by Allcott *et al.* (2020), on US subjects during the four weeks leading up to the 2018 midterm election. The subjects who abstained from Facebook during this time were found to have lower political polarisation, compared to those who used Facebook as normal. (They also had increased subjective wellbeing, and reduced factual knowledge about the news.) A similar study was conducted by Asimovic *et al.* (2021), on subjects from Bosnia and Herzegovina, during the 2019 remembrance week for the Srebrenica genocide. Contrary to what we might expect from the US study, the group who abstained from Facebook showed lower regard for ethnic outgroups than the control group.

Finer-grained studies ask or induce volunteer users to behave in certain ways on social media. For instance, a study by Levy (2021) randomly offered US Facebook users free subscriptions to liberal or conservative news outlets. Levy found that participating users' subsequent visits to news sites were affected by the slant of the outlet they subscribed to, and that users subscribing to an outlet oriented away from initial political attitude reduced their negative attitudes towards the opposing political party. While Levy found changes in political *attitudes*, she did not find any change in political opinion—an interesting mixture of results.

Studies of this kind surface population-level effects, and don't explicitly study the role of recommender systems. They may also suffer from sampling biases in recruitment of participants. More generally, there are limits to how users can be asked to alter their behaviour without behaving unnaturally, which limit the granularity achievable by intervening in user behaviour. But these studies also remind us of the point already made in Section 3.1—that there are likely to be differences in the effect of social media platforms on users, across different countries, and over different times.

3.6 Summary

This chapter has reviewed studies using 'external' methods to measure the effects of social media, and in particular of recommender systems on users. We reviewed five basic methods for studying these effects with-

out access to company-internal data: large-scale population studies (Section 3.1), browser logging studies (Section 3.2), studies using public APIs (Section 3.3), studies using robot users (Section 3.4) and studies intervening in users' behaviour (Section 3.5).

No single story emerges from the studies we reviewed. Some find evidence of sizeable worrying effects: for instance, robot user studies consistently find worrying effects of YouTube's recommender system. Other studies find no evidence of worrying effects of recommender systems: Ledwich and Zaitsev's API study is a case in point. Other studies show evidence for small worrying effects: Flaxman *et al.*'s and Hosseinmardi *et al.*'s logging studies are good examples here. Other studies draw attention to other factors that probably contribute to the worrying effects at issue. Population studies are particularly good at identifying these other factors.

This complex picture is certainly due in part to the fact that effects of recommender systems on users vary as a function of time, and place, and platform. This finding surfaced many times in our review, in studies using different methods. But we don't think the complex picture is only due to this variation. Another issue is that all the experimental paradigms reviewed here have serious *methodological shortcomings*, as we have noted throughout our review.[3] To recap: population studies have systematic problems with confounding variables (as we discussed in connection with Boxell *et al.*'s 2017 and 2020 studies). Browser logging studies raise concerns about sampling volunteers (as we discussed in relation to Flaxman *et al.*'s 2016 and Hosseinmardi *et al.*'s studies). Sampling problems loom particularly large for experiments aiming to study radicalisation effects: users in the process of becoming radicalised are less likely to agree to participate, even if offered money. Hosseinmardi *et al.* also acknowledge the lack of recommender system data in their logging study. Studies using company APIs to study recommender outputs are hampered by the fact that APIs only surface recommender outputs for a 'generic' user account, and entirely fail to address how recommender systems learn about individual users (as we discussed in relation to Ledwich and Zaitsev's 2019 study). Studies of robot users are not studies of actual users: robot users may not interacting with recommended choices in the ways that real people would.[4] Studies that ask actual users to behave in particular ways may also induce unnatural behaviours (as we discussed in relation to Levy's 2021 study). These latter studies also raise concerns about sampling biases.[5]

In addition to these issues, a central problem for all the empirical paradigms reviewed here is that they do not test *causal hypotheses* about the effects of recommender systems on users. Testing a causal hypothesis about the influence of A on B requires actively intervening in A, and looking for effects on B: see Pearl (2009) for a good statement of this principle. If we want to know about the *causal effects* of recommender systems on users, we need to experiment with *changes* to recommender systems, and measure subsequent user behaviour. None of the experiments reviewed here *manipulate the recommender system* to observe its effects on users. The methods reviewed in Sections 3.1–3.3 explicitly surface results about *correlations*, rather than causations. The methods reviewed in Sections 3.4 and 3.5 manipulate real or simulated users, in ways which could be argued to stand in for manipulations of the recommender system—but nonetheless, a paradigm which directly manipulates recommender systems is certainly preferable as a way of evaluating causal hypotheses about their effects.

We want to emphasise these methodological problems, because the methods we will propose for studying recommender systems effectively address all of them: confounding variables, sampling bias, shortcomings due to APIs or approximations of actual user behaviour. In particular, the methods we propose allow genuinely *causal* hypotheses about the effects of recommender systems to be tested, by presenting different versions of the recommender system to different groups of users, and looking for differences in the behaviour of the different groups. The methods we have in mind are exactly those the social media companies use *themselves*, to develop and optimise their recommender systems. We will review these 'company-internal methods' in Chapter 4. In Chapter 5 we will propose a way of leveraging these methods to ask the question at issue for

[3] Our findings about methodological shortcomings echo the findings of UNESCO's broader and more international review of social media impacts on extremism material (Alava *et al.*, 2017). This report, drawing on a larger pool of studies than we surveyed, also found most reviewed studies were 'of low methodological quality', and relied on 'limited data sets'.

[4] Likewise, studies showing problems arising in formal models or simulations of social media systems, of the kind reviewed in Sections 1.3.3 and 1.4, aren't studies of actual social media systems: they establish a 'cause for concern', but nothing more.

[5] Recall that our discussion entirely has omitted studies relying on self-report of social media effects, which have their own methodological issues (as reviewed by Griffioen *et al.*, 2020).

us: whether recommender systems have any role in leading users towards harmful material.

4 A review of 'internal' methods for studying the effects of recommender systems on users

The details of how social media companies study the effects of their recommender systems on platform users are a commercial secret, just like the details about how these systems are trained. But we do know a lot about the different methods that are used. In this chapter we will review what is known about these methods. There is a common distinction made between a set of 'online' methods, that have been used for many years, and a set of 'offline' methods that have been adopted more recently (including the 'bandit' methods briefly discussed in Section 1.1.1). We will discuss online methods in Section 4.1, offline methods in Section 4.2, and 'hybrid' methods combining online and offline techniques in Section 4.3.

4.1 Online methods

4.1.1 Traditional A-B testing

Early user testing of recommender systems employed a technique known as A-B testing (see Shani and Gunawardana, 2011). In this technique the users are split into two groups: A and B. Each group engages with a different recommender system, and the behaviour of the users is observed. The better algorithm will put better results closer to the top of the results list and so the A-B experiment is run simply to determine how far down the result list, on average, a user clicks in order to get the content they desire—a metric known as 'average clickrank'.

Some care must be used to ensure the robustness of A-B testing. The A-B split must result in two demographically equivalent groups. The experiment must be run for long enough to get statistically sound results. And the number of users in each group must be sufficiently large.

Social media companies continue to run A-B experiments. With a user base of billions, it is quick to run an experiment with real users, whereas lab simulations can be prone to effects caused by simulation metrics. Indeed, social media companies run many A-B experiments in parallel: the users who are participating in these experiments are of course entirely unaware that they are doing so.

4.1.2 Interleaving and multileaving

A more subtle and informative way of comparing two different recommender systems is to *combine* their outputs in the feed seen by users, and then draw conclusions about the quality of different items in the lists from user clicks.

Interleaving methods treat the ranking of content items produced by a recommender system as a set of *partial* rankings (see e.g. Joachims, 2002; Chapelle *et al.*, 2012). An analysis into partial rankings is important because users don't process all items in a ranked list equally: instead, they process them serially, roughly in the order they are ranked, until they find something good. If they click first on the item ranked N, then we know they prefer it to all items ranked higher than N. That is, we have learned something about relative preferences. We want to evaluate a recommender system by the set of relative preferences it delivers. Interleaved methods basically define a way of presenting two alternative rankings in a single ranking presented to the user, in a way that keeps the user's experience as similar as possible to the way it is when getting content from a single recommender algorithm. When a user clicks on one item, that comes from ranker A, we can infer a preference for relative rankings for all items delivered by A earlier in the list. If users always click on a result from A and never from B then A is clearly better than B.

There are several variants of the interleaving approach, which are usefully discussed by Chapelle *et al.* (2012). For instance, 'balanced interleaving' (e.g. Joachims, 2002) ensures that the top k results in the final list always contain the top k_a results from recommender A and the top k_b results from B. 'Multileaving' (e.g. Brost *et al.*, 2016) builds the interleaved list iteratively, in rounds. In each round, the recommenders are shuffled and the top from each, in turn, is added to the final list.

4.2 Offline methods

A-B testing and interleaving are both online evaluation methods, where different recommender algorithms are placed before users, and assessed 'in the field'. But there are also offline evaluation methods for evaluating (and developing) recommender algorithms. Offline methods make use of data already gathered, to test different alternative recommender systems. An obvious use for offline evaluation is to sanity check a new algorithm before putting it in front of users.

4.2.1 Methods using supervised machine learning

Some offline methods employ a supervised learning paradigm, of the kind described in Section 1.1.1. The aim here is to predict the user's next click in some pre-existing dataset of user behaviour. Until recently, most academic research about learning recommender systems used this supervised approach.

In its simplest form, a user has a preference for a single document from the collection (the document they will click on). The aim of the supervised learning is to train a scoring algorithm to put that document at the top of the results list. The problem is known as 'known entity search' in the Information Retrieval literature, or as 'information refinding' in the web literature. The document *is* in the collection, there *is* only one of them, the task is to put that document at the top of the list. Of course, that one document might be different for different users, and so a rank ordering of documents based on the number of users who identify each document as 'correct' can be constructed from a click log.

A large enough log of a large enough number of user clicks can serve as the training set for a supervised method of this kind. The quality of the resulting algorithm can be determined using metrics such as nDCG (Järvelin and Kekäläinen, 2002). More thorough evaluation can be done by using the trained algorithm in an A-B experiment. The A-B experiment is performed in order to validate the results of the offline experiment before a site switches over the new algorithm as the default.

4.2.2 Methods using reinforcement learning

A new generation of offline methods using reinforcement learning rather than supervised learning (the 'bandit methods' mentioned in Section 1.1.1) was heralded by a paper by Bottou *et al.* (2013). Bottou *et al.* argue that learning a ranking algorithm is much better modelled as a reinforcement task than as a supervised learning task, because the behaviour of the ranking algorithm *affects the behaviour of the user*. In online paradigms, we can directly deploy rankers and assess the resulting user behaviour in a A-B experiment. But it's hard to search the large space of possible rankers using online methods. Bottou *et al.* devise a method for running simulated experiments with different rankers, and estimating the effects of these experiments using data gathered from a single actual experiment.

Bottou *et al.*'s learning model is essentially a model that learns about a *causal mechanism*, typically modelled as a directed acyclic graph (DAG) whose arcs represent causal interactions (see e.g. Halpern and Pearl, 2005). A standard causal model of a recommender algorithm allows us to compute various performance metrics, by running simulations forwards. (To cater for unknowns, and model the cyclic nature of recommender system causal influences, an approximation is developed in the form of a Bayesian network.) Within this approximate model, Bouttou *et al.* devise a way of estimating the performance of a *counterfactual* recommender system with specified properties. The basic idea is to rerun a simulation with counterfactual input variables. The key technique is to 'reweight' outcome variables. Reweighting involves a process of 'importance sampling', to accommodate the fact that reweighting changes the distributions of variables in the system. Importance

sampling requires that the data from which conditional probabilities are estimated should be gathered from experiments involving 'active randomisation' (i.e. experiments which actively flatten data distributions). The noise added to the experiment helps estimate counterfactual probabilities.

Having devised a way of evaluating the performance of a counterfactual recommender system with specified properties, Bottou *et al.* then use this to define a method for *training* a recommender system to optimise the chosen performance measure. Basically, they explore a *space* of counterfactual recommender systems, and find the properties which make the recommender system perform best by the chosen measure. Importantly, the training method also relies on gathering data from experiments involving active randomisation: the data thus gathered offer 'precious cues' and 'useful signals' for learning algorithms.

There are many variants of reinforcement learning algorithms based on Bottou *et al.*'s counterfactual approach. In the literature, they are often referred to as 'bandit' methods. (The model they build falls within the class of 'contextual bandit' reinforcement learning models, which extend a well known class of 'multi-armed bandit' reinforcement learning models.)

4.2.3 Aside: causal inferential methods and algorithmic transparency

The bandit methods mentioned in the previous section are a special case of a more general suite of techniques which fall under the scope of causal inference methods. Bandits are designed to play in a reinforcement learning setting, which is common in the game theoretic scenario of a social media platform. However, this should not preclude the possibility of an experimental design to tease out causal effects in social media interactions, that uses causal inference in a way different to reinforcement learning.

This may be particularly important when in addition to causal mapping of user interactions using bandit methods, it might also be pertinent to gain causal insight into recommender systems themselves, and into the decision-making processes they implement. As machine learning processes have become more powerful in recent years, the structure of trained systems has become much harder to interpret, and their decision making processes are often very opaque (Chakraborti *et al.* 2020). This is especially true for neural networks, which make up a large component of recommender systems, as we saw in Section 1.1.1. Transparency as 'explainability' is not the main focus of our current report, which focuses on transparency of effects. But we mention this point, because causal models have considerable potential to contribute to explainability too (see e.g. Miller, 2019 for a review).

To spell out this idea: recall from Section 4.2.2 that causal inference relies on a model of the interactions taking place, expressed as a directed acyclic graph (DAG) whose weighted arcs represent the directions, quality and extent of interactions (Halpern and Pearl, 2005). For the social media bandit system just discussed, the nodes of the graph are users, content items, clicks, and so on. To model the decision process of a trained neural network, the nodes could be the neurons in the network, or other components of the trained model. In either case, structural equations can then be drawn up using the DAGs, on which counterfactual logic can be applied. Note that for causal inference to work, we need to have prior knowledge of the appropriate graph model, and the dataset must be large enough that examples of all interaction and decision possibilities would be present to measure the causal flow. This would be a challenge for recommender systems—but perhaps not an insurmountable one.

4.3 Hybrid methods

In practice, it is infeasible to train a recommender algorithm from scratch using live users. Not through any lack of data, but more because of the consequences of putting partially learned (badly performing) algorithms in front of users. The initial state of some of these algorithms is purely random and consequently the user experience would be truly dreadful—imagine being recommended a completely random item on a social media site. It's also suboptimal to train recommender algorithms using purely offline methods: we want our final evaluations to be on actual platform users.

A combination of techniques is therefore needed. In practice, offline tests (especially bandit studies) are used to learn several candidate recommender systems, which are then placed before actual users in online tests (probably using some form of interleaving). The best performing systems in these tests are then used as seeds for a further round of offline tests, and the cycle continues.

4.4 Summary

In Chapter 3 we reviewed methods available to external researchers for studying the effects of recommender systems on platform users, and in the current chapter we reviewed methods available to researchers working within companies. We are now in a position to make some comparisons between the two groups of methods.

The central point we want to make—-perhaps the central point in this whole report—is that the internal methods described in the current chapter are far superior to the methods available to external researchers. Perhaps this is hardly a surprise, given the amount of data available to company-internal researchers, and their direct access to the relevant algorithms—but we feel it is still important to make this contrast. As summarised in Section 3.6, the methods available to external researchers wanting to study possible harmful effects of recommender systems on platform users all have serious methodological shortcomings. Population studies have problems with confounding variables; browser logging studies and user manipulation studies have problems with possible sampling biases; studies using public APIs can only access recommendations made for a 'generic' user, and have no way of gauging the crucial effects of personalised output; studies of robot users aren't studies of actual users.[1] Most critically, none of the external methods reviewed in Chapter 3 test properly *causal* hypotheses about the effects of recommender systems on platform users, because they cannot systematically *manipulate* the recommender system.

The company-internal methods we have described in the current chapter don't suffer from any of these problems. Right off the bat, these methods are all about trying out different recommender systems on users. Online methods (Section 4.1) overtly do exactly this: so they explicitly examine the causal effects of recommender systems on actual users. Offline bandit methods (Section 4.2) try out different 'counterfactual' recommender systems on logged user data, rather than in live experiments—but here too, these models explicitly evaluate causal hypotheses about recommender system effects on users: the whole paradigm involves the learning and deployment of causal models, as discussed in Section 4.2.2. Moreover, none of these methods need suffer from sampling biases. Users are recruited for these experiments without their knowledge (for better or worse); and vast amounts are known about user demographics, so in principle, experimenters should be able to draw fair samples from any given population to run their tests. Furthermore, there are no problems with confounding variables. In company-internal tests, users are randomly assigned to groups exposed to different recommender algorithms: if there are any differences in the behaviour of users across groups, these can be unambiguously attributed to the recommender algorithm. (Note in particular that other parts of the technical system can be held constant over these experiments. For instance, the content moderation mechanisms in force can be held constant, to isolate the effects of the recommender system proper.) Finally, and most obviously, the experiments conducted internally to companies run on real users, on actual social media platforms. And to cap it off, they are informed by vast amounts of data, because they run on the vast user base of social media platforms, and exhaustive data is gathered about every aspect of users's lives and behaviour.

[1] The formal models and computer simulation models we presented in Chapter 1 are also not studies of actual users, or of actual systems.

5 A fact-finding exercise using 'internal' methods, using New Zealand as a case study

The argument of this report so far is as follows. We argue there is a *prima facie* concern about the effects social media recommender systems have on social media platform users (see Section 1.6 and Section 2.4.6). The concern is that recommender systems may have a role in moving users towards various kinds of harmful content, and in particular towards TVEC. It is incumbent on external stakeholders—in particular on governments and civil society groups—to investigate this concern: empirical studies are needed, to ascertain the effects of recommender systems on users. We argue that empirical studies performed *externally* to social media companies cannot answer the question satisfactorily, because the methodologies available all have significant problems (see Section 3.6). But these problems can be overcome in studies conducted *internally* to social media companies (see Section 4.4). Social media companies, of course, also have a pressing interest in understanding the effects of their recommender algorithms in relation to harmful content, as we noted in Sections 1.6 and 3.3.5. As a conclusion, therefore, we recommend a *collaboration* between external stakeholders and companies, in which company-internal methods are used to study the effects of recommender systems on users' relationship with harmful content. We call this collaboration a 'fact-finding exercise'.

In this chapter, we make various suggestions about how the fact-finding exercise could be conducted. We begin with some considerations relevant to policy. In Section 5.1 we suggest an international framework that provides a natural context for the exercise we have in mind (namely the Christchurch Call's 2021 workstream on 'algorithms and user journeys'). And in Section 5.2 we describe work we are currently undertaking in New Zealand, to implement the fact-finding exercise using New Zealand as a case study country. The fact-finding exercise and its relation to the Christchurch Call will be further discussed in Annex C, with a focus on legal and policy issues. In Sections 5.3–5.7 we outline various technical suggestions about how the fact-finding exercise could be conducted. We finish in Sections 5.8 and 5.9 by discussing issues relating to auditing, transparency and privacy.

Our suggestions in Sections 5.3–5.7 should be understood as starting points for a broader discussion with company engineers, which we are keen to pursue. We have already had informal discussions with engineers from several companies, that confirm the basic feasibility of the proposed methods. (Exactly how the fact-finding exercise would play out in a given company will of course depend on the resources and methods it uses—in particular, on the content classifiers that are available.)

Our suggestions in the chapter as a whole should also be regarded as starting points for a broader discussion with companies and external stakeholders about the legal and policy issues raised by the fact-finding exercise. Again, we have conducted our own survey of these issues, which we present in Annex C. Our survey in Annex C catalogues the relevant issues very broadly, considering the perspectives of companies and human rights organisations, as well as the perspective that motivates our report. We hope it will be useful in structuring a broader discussion around our proposed fact-finding exercise.

The New Zealand government has initiated discussions with a selected social media company about our proposed fact-finding exercise, and we have had an initial meeting with company engineers, to discuss the exercise. The meeting was positive, and our discussions will continue.

5.1 An international policy context for the proposed fact-finding exercise

As just noted, social media companies are just as interested as external stakeholders in understanding the effects of recommender systems on users' attitude towards harmful content. This shared interest is most

clearly articulated for TVEC content. Tech companies and governments have already formed a partnership to address the issue of online TVEC, under the umbrella of the Christchurch Call to Eliminate TVEC Online,[1] a collaboration initiated in 2019 following the terrorist attacks on Christchurch Mosques. As already discussed in Section 2.2.1, the initial work of Call participants focused on the definition and identification of TVEC, and on protocols for sharing it and quickly removing it. In this work, the GIFCT and OECD have played important roles, as also noted in Section 2.2.1. But the Christchurch Call also contains a commitment from companies to 'review the operation of algorithms and other processes that may drive users towards and/or amplify terrorist and violent extremist content to better understand possible intervention points'. The review is envisaged as possibly incorporating 'appropriate mechanisms for reporting, designed in a multi-stakeholder process'. In this year's Christchurch Call summit, a programme of work was agreed that advances this agenda:

> As the Call Community we will devote effort and resources towards better understanding the 'user journey' [towards TVEC] and the role this may play in the broader radicalisation process.
>
> We will design a multi-stakeholder process to establish what methods can safely be used and what information is needed—without compromising trade secrets (...)—to allow stakeholders to better understand the outcomes of algorithmic processes and their potential to amplify TVEC.[2]

The workstream as defined here is already being pursued by the GIFCT, as witnessed in particular by its recent report on recommender systems (GIFCT, 2021c). The workstream also provides a perfect context for the fact-finding exercise we have in mind. The fact-finding exercise we outline in this chapter is essentially a suggestion about the 'methods that can safely be used' to allow external stakeholders to better understand the effects of recommender systems on users' relationship to TVEC. Note that these 'methods' should surface the relevant information about recommender system effects without compromising company IP—which is another key factor to consider in designing methods.

There are also several jurisdictions around the world developing legislation about social media oversight. We will review these in more detail in Annex C, but it is worth highlighting the EU's proposals in this area, which envisage very broad powers for social media regulators. The Digital Services Act (European Commission, 2020 art. 64) states the Commission 'may require access to or reporting of specific data' about platform operation, including 'data necessary to assess the risks and possible harms brought about by the platform's systems', and more specifically, 'data on the accuracy, functioning and testing of (...) recommender systems'. These are very broad powers; it is useful to think about exactly what *specific* types of data may be necessary in these contexts. Our collaborative fact-finding exercise can also be understood as an exploration of how existing legislative proposals can be made more precise.

5.2 Our current work: a fact-finding exercise centred on New Zealand

The role of GPAI is to offer technical advice on AI issues and AI methods to its member governments. It is not our role to advise on policy—other organisations (in particular the OECD) have that role. Consequently, we present our proposed fact-finding exercise as the method we advise a government uses to engage with a social media company, *if* that government wishes to find out more about the effects of the company's recommender systems on its citizens. The fact-finding exercise we describe in this chapter is one which we envisage *any* government could perform, in collaboration with *any* social media company, to study the local effects of the company's recommender system.

We construe the fact-finding exercise in this way partly to keep our advice focused on technical methods, according to GPAI's brief. But a model focusing on individual country initiatives also advances an interesting model of *local governance* of social media platforms. As noted several times in this report (see e.g. Sections 3.1 and 3.6), the effects of social media platforms vary considerably by place, and by platform: it makes sense for mechanisms that oversee these effects to have some degree of local scope, so as to be sensitive to these

[1] https://www.christchurchcall.com/

[2] Text taken from the Christchurch Call 2021 workstream document.

variations. We believe it may also make the oversight process more manageable, because it can be trialled incrementally, on a country-by-country basis. The global reach of social media platforms makes them hard to oversee in many ways, but in some cases, country-level governance mechanisms may provide a manageable way of decomposing the problem. (Note that our community consultation project on hate speech presupposes a similar model of local governance, as discussed in Annex A.)

Another benefit of a local governance model is that we can define the *conditions* under which a government can legitimately conduct a fact-finding exercise. We can define the *objective* of the exercise, to ensure this addresses the wellbeing of citizens, to exclude governments pursuing it out of self-interest. We can also define certain regulatory structures within which the exercise is to be conducted. Annex C discusses legal and policy issues raised by our fact-finding exercise; in that Annex we will consider in more detail how the objective of the exercise can be framed, and what regulatory structures could be adopted.

As an initial case study of a country-centred fact-finding exercise, we are working with the government of Aotearoa New Zealand, in collaboration with a selected social media company, to explore the methods outlined in the remainder of this chapter. The government has had initial discussions with the company, and set up an initial meeting between our group and company engineers, to discuss the methods we propose. The meeting went well, and further technical meetings are planned.

Note our description of these methods assumes the fact-finding exercise will study the effect of recommender systems on users' relationship towards TVEC (given the framework of the Christchurch Call), and will focus on New Zealand users (our case study country). But in principle, it could be used to study users' relationship towards any kind of harmful content.

5.3 The basic form of the study

The study we have in mind could employ online methods (Section 4.1) or offline, 'bandit' methods (Section 4.2.2) for observing the effects of recommender systems on users. In either case, we envisage our study will be incorporated within existing studies of recommender systems that are already being conducted by the company to test effects on New Zealand users: we don't envisage any change to the experience of users.

In the case of online methods, the study would randomly identify a number of groups of New Zealand users, and assign a different version of the recommender system to each group. We assume companies already conduct online studies of this kind; we envisage our fact-finding study being incorporated into existing online studies. Incorporating our study would simply involve recording *additional measures of user behaviour* across all groups, assessing users' relationship with TVEC, to examine if different versions of the recommender system impact on this relationship.

In the case of offline 'bandit' methods, we envisage our study being incorporated into existing bandit optimisation methods trained on data from New Zealand users. Here the idea is again to record *additional measures of user behaviour*, assessing users' relationship with TVEC, this time during offline runs of 'counterfactual' recommender systems (see Section 4.2.2). We can then examine whether there is any relation between the user behaviour measures being used to optimise the system and these TVEC-related measures. For instance, we can ask if different 'counterfactual' recommender systems trialled during optimisation have different effects on users' relationship with TVEC.

Note that both online methods and bandit methods test explicitly causal hypotheses about the effect of recommender systems on users' relationship with TVEC. For online methods, the hypothesis is tested by explicitly manipulating the recommender system placed given to different user groups. For bandit methods, we test a causal hypothesis expressed within a causal model built with data gathered from actual users. Online methods and bandit methods also both avoid problems of confounding variables, and of sampling bias, by the reasoning outlined in Section 4.4.

The crucial question to address, of course, is what metrics should be used to measure 'users' relationship with TVEC'. This is what we will discuss in Sections 5.4–5.7. We reiterate that our proposals here should be

understood as starting points for discussions with company engineers.

5.4 Two types of TVEC-related metric

There are two possible types of metric to consider. **Endpoint metrics** chart behaviours that reference 'actual TVEC' in some way. **Pathway metrics** reference behaviours that are understood as being involved in the processes that lead to radicalisation—that is, the kinds of behaviour we discussed in Sections 2.3 and 2.4).

It's useful to compare endpoint and pathway metrics both as regards their *validity* as measures of a user's 'relation to TVEC', and as regards their *statistical power* in studies of recommender systems.

Validity Endpoint metrics have a high validity, because they measure direct engagement with TVEC of one kind or another. Pathway metrics have a much lower validity, for the obvious reason that users showing signs of being on a 'path towards TVEC' could at any time stop moving along this path. Some analyses of radicalisation certainly emphasise that people's options become progressively more limited as the journey proceeds (see in particular Moghaddam's account in Section 2.3.1); but the process is obviously very stochastic, so no strong inferences can be made about future points on a path.

Statistical power While endpoint metrics have a high validity, they have a low statistical power, because only very small numbers of people engage with actual TVEC. Actually, in some groups, and in some contexts, a surprisingly high proportion of users report searching for TVEC: for instance, 36% of the young adults in Frissen's (2021) study reported having searched for ISIS beheading videos. But if we sample from the general population, endpoint metrics will certainly identify small numbers of people. We expect the differences in user behaviours due to recommender algorithm to be very small, especially for online methods; if the behaviours of relevance are only exhibited by very small groups, they might be missed. Pathway metrics have higher statistical power, because many more people exhibit pathway behaviours (though the numbers naturally decrease as we move along the pathway).

We should note that the Christchurch Call is interested in pathways in their own right, as well as in endpoints. This year's workstream in particular aims to work towards 'better understanding the 'user journey'' towards TVEC, as noted in Section 5.1.

5.5 Endpoint metrics

Measuring endpoint metrics is somewhat hampered by the fact that TVEC must be removed whenever it is identified. But there are nonetheless ways of measuring behaviours that 'reference actual TVEC' without delivering it to users online.

5.5.1 Number of searches for TVEC

One approach is to measure the number of times a user *searches* for TVEC. Such searches could be identified in different ways, that we won't consider here; the crucial point is that TVEC doesn't need to be returned in order to identify a user's interest in it. There are several initiatives that 'redirect' searches for TVEC or other harmful content; the method used by Moonshot group has been trialled on many social media platforms (see Ganesh, 2020 for a review of this and other methods), so this is likely to be a practical approach.

5.5.2 Engagement with actual TVEC

Engagement could involve production of new TVEC items (posting, sharing) or consumption (viewing) of existing items. TVEC which is posted or shared will of course be blocked, if it is identified, but the platform

can still count these events. Counts will be very small, because users posting TVEC items are normally deplatformed, but even small numbers have high validity, as already noted.

As regards consumption, it is obviously not possible to leave TVEC on the platform and wait for users to find it. However, consumption can be measured in stored logs of user behaviour, for TVEC that is only identified in retrospect. If a platform kept detailed enough logs during online A-B tests in the past, the number of times users viewed actual TVEC is an endpoint metric that could be used to look for effects retrospectively. (We have reason to believe that platforms do keep logs of the right kind.) Note that this metric can also be used in offline bandit experiments—but again, only on TVEC that is identified retrospectively, after the training data for the bandit experiment was gathered.

5.6 Pathway metrics

5.6.1 Engagement with 'TVEC-adjacent' content

Many platforms identify content that just falls short of TVEC, that is not removed, but demoted in some way. For instance, YouTube and Facebook both define several categories of 'borderline' harmful content that receive reduced distribution and/or are flagged in some way (see Section B.2 for details). One possible pathway metric is a measurement of users' engagement with borderline content of this kind. We should emphasise that what counts as borderline harmful content is often quite distant from actual TVEC—but some categories are closer than others. For instance, Facebook has a category of 'content posted by groups and Pages associated with (but not representing) violence-inducing conspiracy networks, such as QAnon'. If it were found that different recommender algorithms cause different amounts of engagement with content of this kind, that would be an interesting result.

We expect to see more users engaging with TVEC-adjacent content than with actual TVEC—partly because the numbers of people engaging with content diminish as it becomes more extreme, but also because TVEC-adjacent content is not removed. This fact also makes online testing protocols more straightforward. But we are still faced with the issue of validity: if our main concern is with people engaging with actual TVEC, what is the significance of a metric measuring engagement with content that is close to TVEC, but *not* TVEC?

5.6.2 Behaviours measured on a continuum of 'TVEC-relatedness'

Another type of pathway metric can be derived by rating content on a continuous scale measuring 'relatedness to TVEC'. There are various possible methods to consider here. One method requires a 'general' content classifier, that rates every content item directly on a scale from 0 to 1, where 0 denotes content that is unrelated to TVEC and 1 denotes actual TVEC. Other methods require more specific classifiers, that rate content using continuous measures that have been hypothesised to play a role in radicalisation, surfacing for instance the average 'intensity of hate' in the content a user engages with (Section 2.2.3), or the average amount of moral emotion (Section 2.4.1) or moral outrage (Section 2.4.2), or the average proportion of items referring to political out-groups (Section 2.4.3), or containing falsehoods (Section 2.4.4). The amount of polarisation is another interesting pathway metric, but note this applies to a group, rather than to individuals; this may make it hard to use in online trials, since the relevant groups may cut across experimental groups. But polarisation is a possible pathway measure in bandit experiments, where there is no division of users into groups. The bounds on these more specific metrics are harder to define, but they nonetheless all rate content on a scale that varies continuously.

If we can rate users' relationship to TVEC on some continuous scale, then we can compute averages and standard deviations on this scale for groups using different recommender systems, both in online experiments and in bandit experiments. We could then look for differences between groups at a single point in time. In addition, we could compute gradients over time for each group, denoting how much the group has changed in its relationship to TVEC. We could also look for differences in gradients across groups using different

recommender algorithms.[3] On all these measures, we expect to have good statistical power, because there will be variance over the whole group of users: we are not just measuring the incidence of rare behaviours only found in extremists. But again we are faced with questions about validity. Say we define a user's 'TVEC rating' during a certain period as the average rating of content engaged with by that user on the 0-1 scale of TVEC-relatedness mentioned above. And say when we trial two recommender systems on two user groups, we find an average TVEC rating of 0.1 for one group and 0.2 for the other. Can we conclude anything useful about these recommender systems if our main interest is in the effect of recommender systems on users' engagement with *actual TVEC* on the platform?

5.7 Validation mechanisms for pathway metrics

In Section 5.6, we introduced two types of pathway metric, that measure users' relation to TVEC 'at a distance'. TVEC-adjacent metrics measure behaviours of users whose distance to TVEC is minimal; 'continuum' metrics measure behaviours at all possible distances. There are important validity uses for both types of metric: it's not clear what they tell us about the endpoint of a journey towards TVEC. In this section, we will consider two ways we might look for validation.

5.7.1 Mechanisms using predictive models of radicalisation

One thing we could do is to build a model that *quantifies* how much evidence 'pathway' measurements furnish about the likelihood of full radicalisation. Internally to social media companies, predictive models can readily be learned, using logs of user behaviour. Companies keep extensive logs—the details of how logging happens are again not public, but if logging is extensive enough, and includes logs of users who became radicalised enough to interact with actual TVEC, then a predictive model of a process leading to this level of radicalisation could in principle be learned.

To do this, we would have to provide an operational definition of a user who is 'fully radicalised', in terms of online behaviours: in our terminology, some definition given in terms of 'endpoint metrics' would be a natural one. We could then use one of several statistical paradigms to learn a predictive model. To be concrete, we will spell out how this could work using a method that is commonly used to predict events that can occur at unspecified points in the future: Cox regression. This method is widely used in medical contexts, to predict outcomes of medical treatments and compare different treatments (see e.g. Benitez-Parejo *et al.*, 2011), and in criminal justice contexts, to predict reoffending rates, and compare different rehabilitation regimes (see e.g. McNiel and Binder, 2007).

A Cox regression model, also called a 'survival analysis' model, predicts the rate at which the outcome event happens per unit of time, as a function of observable variables. In our case, the 'outcome event' would be a user becoming 'radicalised' (i.e. as having some predetermined level of engagement with TVEC content). The observable variables would be the user's behaviours, as assessed by pathway and endpoint metrics. A survival analysis model constructs a 'survival graph', which charts the probability of the outcome event not yet having happened, for each point in the future: probabilities progressively decrease as time advances. In our case, the survival graph would chart the decreasing probability of avoiding 'radicalisation' over time, for users exposed to a given recommender system. The critical variable is the area under the survival graph, which can vary for users exposed to different recommender systems.

This kind of model would provide a quantitative supplement to studies like that of Baugut and Neumann (2020, Section 2.3.7), that ask radicalised people to trace their history of engagement with social media. To the extent that pathway metrics measure features of the content produced by users, this kind of predictive model could also be thought of as quantifying the degree to which users' speech is 'dangerous', in the sense defined by Susan Benesch and colleagues (Benesch *et al.*, 2021). Crucially, with quantitative measures of the kind just described, we can provide a meaningful interpretation of differences in the effects of recommender

[3] Continuous scales of harmful content have advantages relating to annotation protocols, as we noted in Section 2.2.3: they allow comparison-based protocols for annotators, and they allow disagreements between annotators to be resolved using averaging (see again Aroyo *et al.*, 2019; Kritchenko and Nejadgholi, 2020).

systems measured using pathway metrics. Again, the relevant measures can only be computed using data that comes from within companies, because they rely on logged platform behaviours that can only be made by platforms.

We want to emphasise that the role we envisage for predictive models is in validating pathway metrics, rather than in actually making predictions about individual platform users. As discussed in more detail in Annex C, predictive models can be used in surveillance of individual users: in this role, they raise a host of different questions, relating to user privacy and other rights. Whether predictive models can be built for our validation purposes without raising these other questions is a matter for further thought, both at the technical level and at the legal level.

5.7.2 Mechanisms using models of 'homogeneous processes' in radicalisation

We want to finish with an idea about validation that connects particularly to the models of radicalisation and of recommender system effects discussed in this report. The idea relates to the hypothesis about radicalisation we first raised in Section 2.3.5: namely, that some components of the radicalisation process might be 'homogeneous', in the sense of operating in the same way at each point in the process. The mechanisms we had in mind related to formation of increasingly narrow in- and out-groups (van Stekelenburg, 2014), progressive narrowing of behaviours and options (Moghaddam, 2005), progressive heightening of emotions (Brady *et al.*, 2021), progressive consolidation of 'myths' (Moghaddam, 2005). In Section 2.3.5 we also raised the possibility that recommender systems may also be homogeneous in their effects on users during the radicalisation journey—but as we noted at the time, this is an empirical question.

If we are working inside a platform, we have the quantitative tools to test this question. In fact, the 'continuous' pathway metrics introduced in Section 5.6.2 allow hypotheses about homogeneity to be readily stated. For instance, consider the 'TVEC-rating' metric, which measures the average content a user engages with during some period on a scale varying from 0 (unrelated to TVEC) to 1 (TVEC). The hypothesis that a given recommender system operates homogeneously in relation to this scale states that if some amount of exposure E to this system moves users from a rating of x to a rating of $x+n$ on the scale, the same amount of exposure will move those same users from a rating of $x+k$ to $x+n+k$ for all k[4] This is a hypothesis that can readily be tested, by consulting data from users at different points x on the scale. Crucially, if there is some evidence for homogeneity at points midway along the TVEC scale, this provides some validation of the TVEC-rating pathway metric as a measure of something meaningful in relation to 'actual TVEC'. If a recommender system moves users from x to $x+n$ at several midway points on the TVEC-rating scale, we expect by induction that it will move users from $1-n$ to 1 on this scale.

Note that if we have evidence that recommender systems have homogeneous effects, we also have something particularly meaningful to say about *differences* between recommender systems. If recommender system A systematically moves users from x to $x+k_1$ for all x on some TVEC-related scale, and recommender system B moves users from x to $x+k_2$, where $k_2 > k_1$, we have some reason to be more concerned about system B, and some reason to prefer system A, even if we our ultimate concern is only with users' engagement with 'actual TVEC'.

5.8 How should the fact-finding exercise be organised?

In Sections 5.3–5.7 we made some technical suggestions about the form a fact-finding exercise conducted within a social media company could take. In this section we consider how the exercise should be organised, and how its results should be communicated to external stakeholders.

We have already noted that the technical form of the fact-finding study should be a matter for negotiation between external stakeholders and the company in question. Company engineers are clearly the ones with expertise about the mechanisms available within the platform, on all the issues raised in Sections 5.3–5.7. Our main proposal is that the study should be co-designed in such a way that its method and results can

[4] For all k keeping the rating in the range 0-1, naturally.

be published, in the form of a scientific paper. Exactly how the paper describes its methods and presents its results can be negotiated in advance, as part of the co-design process. It's good practice for scientific studies to be designed to produce results in a specified format, so deciding on this format in advance is a conventional part of the design process.

Clearly, the study must be designed so it doesn't compromise the company's IP, or the privacy of platform users. We will consider those issues in Section 5.9. In the current section, we will first outline why we recommend surfacing the results of the fact-finding study in the form of an academic paper. We will then make a practical proposal about how the study could be conducted.

5.8.1 Why present the fact-finding study's results as an academic paper?

A key motivation for the fact finding exercise is to surface *useful* and *reliable* information about the effects of recommender systems to external stakeholders, and to the public. An academic paper would contribute to the wider scientific discussion about recommender systems and their effects that we have reviewed in Chapter 3. It's useful to contrast this format of communication with the ways the public currently finds out about company-internal studies of recommender systems.

We are fairly confident there have already been company-internal studies of the effects of recommender systems on metrics relating to harmful content. These studies may even have used the methods we have proposed in this chapter. But for the general public, the studies that have been conducted so far are shrouded in mystery. What we know about them has come from articles by journalists, reporting off-the-record discussions with anonymous company employees, or through the revelations of whistleblowers. For instance, Karen Hao in her MIT Technology Review article quotes anonymous ex-Facebook employees as saying that 'study after study' confirmed 'models that maximize engagement increase polarization' (Hao, 2021). More recently, the whistleblower Frances Haugen referred to private Facebook studies showing that engaging content 'inspires people to anger', and that a 'safer' recommender algorithm 'makes less money'. We applaud these initiatives—but the manner in which information about company studies is surfaced is highly unsatisfactory, as we are sure Hao and Haugen would agree. We end up with scandal-ridden newspaper headlines about findings that are stated in the vaguest possible terms, and that can't be substantiated. This is as unhelpful for companies as it is for external stakeholders. We believe it would be far better for companies and stakeholders to co-design a fact-finding study, along with a protocol for communicating its design, and its findings, so that the relevant information can be emerge clearly and quantitatively.[5]

Publicly accessible scientific papers will also contribute to a wider scientific understanding of the effects of recommender systems. At present there is virtually no public science relating to 'company-internal studies' of recommender system effects. Companies probably perform their own private studies, but progress in understanding would certainly be faster if there was communication about methods and results. This would be particularly helpful for the smaller companies, which lack resources to conduct their own research programmes. One of the recommendations of the GIFCT's recent report on recommender algorithms (GIFCT, 2021c) is for increased cooperation between platforms and the scientific community. The report concluded that 'finding new ways for platforms to share data with the research community will be critical to improving our understanding of algorithmic outcomes'. We concur—and we further suggest that this sharing process can be helpfully advanced by published papers reporting on fact-finding experiments of the kind discussed in this chapter.

[5]We should note that Frances Haugen has lodged a large number of documents with the US Securities and Exchange Commission: it may be that some of these bear on recommender system studies, and supply some of the detail needed. But this one-off disclosure certainly wouldn't remove the need for *general methods* allowing external stakeholders to ask social media companies quantitative questions about the effects of recommender systems. We need methods that can be implemented for different companies, and at different times. (*Regular* monitoring of recommender system effects is important: as discussed in Section 3.1, effects are likely to vary over time, as well as over place.)

5.8.2 Auditing processes, and a practical proposal

We suggest the most practical way of conducting the fact-finding exercise is to place an external researcher within the company's engineering team, operating with the guidance of that team, under a non-disclosure agreement. We understand that academic research groups collaborating with social media companies often 'embed' researchers in this way. The researcher will, of course, learn more about company-internal processes than will eventually be surfaced in the paper describing the fact-finding study. The non-disclosure agreement will prevent the researcher from communicating anything about this additional learning. However, it will allow the researcher to function as an *auditor* for the reported study, to vouch that it was conducted as described.

It is important that there should be an auditing process for any fact-finding exercise that is conducted. The external auditor need not be an embedded researcher, of course—there are other models where the fact-finding exercise is implemented by company engineers, and the external auditor is not actively involved. But an auditor who is actively involved in the exercise is certainly in the best position to vouch it was done as reported.

5.9 Safety of the proposed fact-finding exercise

It is essential that a fact-finding exercise conducted within a company doesn't endanger company IP, or the privacy of user data. In this section we argue that the form of the exercise we have recommended in this chapter will be safe, in both these respects.

5.9.1 Guarantees for company IP

Our proposed fact-finding exercise provides a measure of transparency about the *effects* of company recommender algorithms—but importantly, it discloses nothing about the *content* of these algorithms. The report we envisage would mention metrics of user behaviour, of the kind suggested in Sections 5.5 and 5.6—but these measures don't relate in any direct way to the inputs and outputs of recommender systems. The fact-finding study doesn't require any disclosure about the internal design of recommender systems, even the broad strokes that are already in the public domain, of the kind outlined in Chapter 1. A class of transparency methods called 'black-box' methods allow access to an algorithm through its inputs and outputs (see GIFCT, 2021c). Our proposed study doesn't even require access to these; it simply observes the behaviours of users.

There are other models of transparency for recommender systems that require some reference to the algorithm that has been learned for a particular user. For instance, mechanisms that answer the question 'Why was I recommended this?' have that form. We believe there is some utility to these models of transparency. (In particular, we believe that it is useful for external stakeholders to know something about the objective function used to optimise a given recommender system.) But the form of transparency we are advocating in the current report doesn't relate to algorithmic transparency. In relation to taxonomies of AI transparency (see e.g. Gavaghan *et al.*, 2020), transparency about effects is to do with issues of *responsibility* rather than issues of inspectability.

5.9.2 Guarantees for platform user data

Our proposed study also preserves the privacy of user data on the platform being observed. The data the study reports about user behaviours is doubly removed from the behaviours of individual users. For one thing, we report about user behaviour through metrics that abstract away from almost all the substance of the content items that users engage with. All that is surfaced about a particular story or video are metrics like '0.2 on the TVEC scale', or 'contains two moral-emotional expressions'. For another thing, we only envisage reporting measures taken over large groups of users. Nothing about the data or behaviour about individual users will be surfaced in the methods we envisage.

5.10 Summary

In this chapter, we have made some practical recommendations about how social media companies and external stakeholders can collaborate to study the effects of recommender systems. These recommendations fit naturally within existing initiatives for research and oversight mechanisms for recommender systems—in particular, this year's Christchurch Call workstream and the ongoing work of the GIFCT (Section 5.1). They are expressed at a manageable scale, in the form of a case study of effects in a single country (Section 5.2). As starting points for discussion, we suggest technical methods the collaborative study could use (Sections 5.3–5.7), and a format for disseminating its results (Section 5.8). We argue these methods are safe, in relation both to company IP and to user privacy (Section 5.9). Again, we are looking forward to direct discussions with companies on all these matters.

A A community-based method for defining harmful online content

A.1 Introduction

Earth – our shared home – is at a tipping point: the Covid-19 pandemic has made clear to many what has been true for some time, that the critical crises of our time necessitate collective action towards common goals, best represented by the United Nations Sustainable Development Goals (SDGs). The crises we face are diverse: the climate emergency and its immediate impacts on the most vulnerable; the role of the Internet and its technologies is both sustaining and countering violent extremism and misinformation; the challenges of rapidly worsening environmental degradation; and the effects of social and economic inequality on human health and wellbeing. These complex, systemic issues require compassionate, collective decision-making by individuals, whānau (families), communities, civil society, governments, intergovernmental organisations, and policy-makers. These issues, the questions they raise, and their impacts are a critical example of the complex inter-relations and connectivity between and in human society, individual rights and freedoms, the environment, and productivity. Understanding these complex systems is the essential demand of our time.

The Global Partnership on Artificial Intelligence (GPAI) represents a global systematic approach to respond to the role that the Internet and its technologies plays in these critical crises. With a shared commitment to the OECD Recommendation on Artificial Intelligence, and a focus on supporting work towards the realisation of the UN SDGs, this collective approach which includes expertise from academia, industry, government, and civil society is the kind of model of multi-stakeholderism which supports international cooperation and the development of universal principles for best practise. GPAI's mission is to support and guide the responsible adoption of AI that is grounded in human rights, inclusion, diversity, innovation, economic growth, and societal benefit, while seeking to address the UN SDGs.

Focusing on work that shifts towards SDG 16, Peace, Justice, and Strong Institutions and SDG 5, Gender Equality, in 2021, the Responsible AI Working Group funded work on a pilot study, Responsible AI for Social Media Governance, which sought to examine some features of possible collective approaches to address social media-based 'dangerous speech'[1], mis- and disinformation, and hateful or violent expression. We asked how might GPAI facilitate the development of principles and techniques which enable governments and communities to understand and investigate harms associated with social media without infringing peoples' rights to privacy, freedom of expression, and fair democratic processes.

A.2 Responsible AI for Social Media Governance – community consultation

The world's large social media companies exert an increasing influence on the way information and knowledge flows in society. While this influence can be positive, there is growing consensus it is also harmful, serving to propagate mis- and disinformation, extremism, violence and many forms of harassment and abuse. There is also a growing consensus that governments should be more involved in mediating, monitoring, and, where appropriate, regulating the influence of social media companies on the dynamics of public discourse, so these processes are undertaken democratically and systematically, rather than unilaterally by private companies.

Considerable international attention is already being paid to the question of how companies should define, identify and remove terrorist and violent extremist content (TVEC). But there are also concerns that the

[1] Dangerous speech is any form of expression (e.g. speech, text or images) that can increase the risk that its audiences will condone or participate in violence against members of another group. See https://dangerousspeech.org/guide/

recommender systems which disseminate content on social media may have a role in in moving users towards violent or extremist content. Our project concerns itself with these interlinked issues – online harassment and abuse, dangerous speech, and hateful or violent expression, and the role of recommender systems and social media company community guidelines in amplifying or mitigating this harmful content.

Social media harms disproportionately affect women, minoritized communities, faith-based communities, indigenous peoples, LGBTQIA+ communities, and people of colour (D'Ignazio, 2020). These communities are also underrepresented in decision-making about social media, from companies to governments to civil society organizations.

What is needed is are processes by which ongoing, trust-based relationships between governments, civil society organizations and community groups and social media companies, can be developed in order to facilitate inclusive, responsive, and community-centred protocols which establish how harm is experienced by those who are most often the target for dangerous speech, harassment and abuse, how community guidelines work or do not work to enable communities to mediate, moderate, and report harms, and how classification of expression captures or does not capture localised nuances and meanings.

In order to address this problem, we needed to imagine a process of consensus-building within which those voices most often left unheard, and who often come from communities who are most likely to be targeted or scapegoated by dangerous speech, harassment and abuse, could be central to the design, creation, and delivery of proposed solutions. Given that this project aims to be able to propose a set of principles and techniques which could be utilised by governments, communities, and social media companies to embark on collective work to understand and improve recommender systems and other social media platform technologies, we started with a series of key questions: who should be involved in critical stakeholder collectives? How will marginalised voices be prioritised? What questions should be asked – and with whom will those questions and proposed solutions be co-designed and co-created?

Couched in terms of exploring the complex interrelations between social media as a tool for information and knowledge flows, while also being a space within which harmful use or misuse is producing and amplifying mis- and disinformation, dangerous speech and many forms of harassment and abuse, as well as violent or hateful extremism, we developed an approach to enable genuine collective engagement between social media companies, communities, and government. Consensus-building requires localised, situated solutions to universal problems, so this project focused on developing a pilot process for community consultation within Aotearoa New Zealand, with the aim to develop a set of inclusive, responsive and iterative, survivor-led and community centred protocols for classifying harm experienced on social media platforms. This approach is grounded in the Universal Declaration of Human Rights, but also within foundational legal, civil, and social frameworks that are particular to Aotearoa New Zealand, critically Te Tiriti O Waitangi[2].

In establishing locally appropriate norms within an international human rights-focused framework, we hope to contribute to principles for growing international consensus, enabling social democracies to act to build social cohesion in ways that are situated, locally grounded, respectful of rights and responsibilities, and operate within existing agreed principles of international law and regulation.

A.3 Context

The deterioration of democracy and its leadership globally has significant ramifications for civil societies, challenging our understandings of truth, transparency, and justice, particularly the ways in which online discourses and behaviours impact public and private spheres. In Aotearoa New Zealand, the Christchurch Mosque terrorist attack of March 15, 2019 which claimed the lives of 51 Muslim New Zealanders at prayer, provided renewed and horrific evidence that white supremacist ideologies and violent expression of those ideologies are on the rise, and thriving in Aotearoa New Zealand.

[2]Te Tiriti O Waitangi refers to the Māori version of The Treaty of Waitangi, see https://www.waitangitribunal.govt.nz/treaty-of-waitangi/meaning-of-the-treaty/

The Covid-19 pandemic which started impacting Aotearoa New Zealand in February 2020, has seen an increase in Internet use, and a related increase in dangerous speech, online harassment and abuse, hateful or violent expression, and online harm. In a working paper published in August 2020, Soar et al. described how the infodemic[3] was manifesting in Aotearoa New Zealand, and globally:

Aotearoa New Zealand's communities have differential experiences of past pandemics, different measures of health and wellbeing, and different experiences of state services and state intervention. The pandemic and infodemic are also taking place within different nation-states, with different political systems, worldviews, and approaches to healthcare and the role of government. These contexts necessarily inform community and individual responses to the overabundance of information experienced. Understanding how the infodemic has presented in Aotearoa New Zealand enables us to better evaluate ways in which unreliable and untrustworthy information differentially impacts our communities.

The Edge of the Infodemic: Challenging Misinformation in Aotearoa (Talbot and Alaili, 2021) describes how exposure to misinformation is common, and concern about it is widespread, with 82% of respondents concerned about the spread of misinformation in Aotearoa New Zealand. Most New Zealanders think that the Internet is playing a key role, with those surveyed stating that social media users and companies "often spread false or misleading information intentionally". At the same time, 79% of New Zealanders get news or information from social media. Expanding from false and misleading information to categories of dangerous speech, harassment and abuse, and hateful or violent expression, Netsafe New Zealand reported in October 2021 that the July-September 2021 quarter, which coincided with a Delta variant outbreak and lockdown for Tāmaki Makaurau Auckland, New Zealand's largest city, was the busiest period in their history (Walters, 2021). Netsafe New Zealand is Aotearoa New Zealand's independent not-for-profit online safety organization. Recent reporting commissioned by the Department of Internal Affairs, Understanding New Zealand's Extremist Ecosystem (2021) showed Aotearoa new Zealand is experiencing similar patterns of rising hateful and violent expression to other countries, with an increase in far-right and white supremacist or white identitarian ideologies. Many New Zealanders were horrified to learn that per capita New Zealanders post more dangerous speech and hateful or violent expression than Australians or those in the UK (Comerford, 2021).

Aotearoa New Zealand's communities experience and engage with mis- and disinformation, dangerous speech, harassment and abuse, and hateful or violent expression in different ways. Māori scholar and human rights defender Tina Ngata (2020) has described how far-right conspiratorial ideologies are being re-packaged to appeal to Māori:

[this] *illustrates a politicising of the distrust of state, media and science that warrants deep discussion. But what concerns me more is the outright manipulation of our people for an agenda that ultimately exploits the marginalized. It's a strategy that's being orchestrated and advanced by the alt-right, and it deliberately places brown, black, and poor people at its forefront.*

Ko tō Tātou Kāinga Tēnei:[4] The Report of the Royal Commission of Inquiry into the terrorist attack on Christchurch masjidain on 15 March 2019 describes, in chapter 4 'What Communities told us about the broader context within which the terrorist attack occurred', how the Muslim Community Reference Group told the Royal Commission that "different ethnic groups need to be included in this, they are also targets of racism. The report goes on to describe how other religious communities, ethnic communities, and Māori reported experiences of racism, adversity, and injustice. Communities consulted in the process of the inquiry stated that the terrorist attack took place in "a context of widespread racism, discrimination and Islamophobia, where pre-judgements or hostile behaviours (including hate-based threats and attacks) are rarely recorded, analysed or acted on. The role of the Internet in both radicalising the terrorist and enabling the distribution of his livestreamed manifesto with effects that still linger in 2021 in myriad ways across platforms and products in the social media ecosystem is a key reason Aotearoa New Zealand stands out as a case study location for the development of new, collective and survivor-led processes to ameliorate our digital public spheres. Aotearoa New Zealand's communities told

[3]The World Health Organization describes the infodemic as an over-abundance of information – some accurate and some not – that makes it hard for people to find trustworthy sources and reliable guidance when they need it. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200202-sitrep-13-ncov-v3.pdf?sfvrsn=195f4010_6

[4]Ko tō Tātou Kāinga Tēnei: Report of the Royal Commission of Inquiry into the terrorist attack on Christchurch masjidain no 15 March 2019. Vol 1, parts 1-3, chapter 4; (2020)

the Royal Commission of Inquiry that government agencies need to meaningfully and respectfully engage with a broad variety of people (including for example, women, youth, people with a range of national origins) and groups, as opposed to representative bodies or groups.

A.4 The Disinformation Project, Te Pūnaha Matatini

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

Our novel approach enables quantitative measures of volume of inaccurate content, amplification of mis- and disinformation by groups and individuals, tracking of narratives across digital and physical artefacts, geographic spread and means of distribution methodologies, and, significantly, sees qualitative evaluation of the situated nature of the content measured and analyzed. In doing so, we produce critically grounded disinformation studies for Aotearoa New Zealand. Data and analysis are presented in ways which are immediately usable for decision-makers, and media commentary on the harms disinformation and dangerous speech present to social cohesion, freedom of expression, inclusion, and safety.

A mixed methods, complex systems approach is presented.

A.5 Complex systems

Complexity is seen in many fields, but can be characterised by the emergence of qualities in a system that cannot be reduced into simpler elements. The environmental scientist and systems theorist Donella Meadows (2008) explained that a complex system is something that is more than the sum of its parts. Philip W. Anderson (1972), the theoretical physicist who described symmetry breaking, showed that at each level of complexity novel properties appear, and that understanding these new features demands new research and methodologies. Intersectionality, first described by Kimberlé Crenshaw (1989), the Black feminist scholar, reveals how layered identities exacerbate discrimination. Māori educational theorist Linda Tuhiwai Smith, in her ground-breaking work, *Decolonizing Methodologies* (2012) writes of the inadequacies of reductionist approaches:

> ...imperialism and colonialism brought complete disorder to colonized peoples, disconnecting them from their histories, their landscapes, their languages, their social relations and their own ways of thinking, feeling and interacting with the world. It was a process of systematic fragmentation which can still be seen in the disciplinary carve-up of the indigenous world: bones, mummies and skulls to the museums, art work to private collectors, languages to linguistics, 'customs' to anthropologists, beliefs and behaviours to psychologists. To discover how fragmented this process was one needs only to stand in a museum, a library, a bookshop, and ask where indigenous peoples are located. Fragmentation is not a phenomenon of postmodernism as many might claim. For indigenous peoples fragmentation had been the consequence of imperialism.

Emergent properties – be they imperialism, broken symmetries, ecosystems, or intersectionalities – require approaches that regard the constituent parts of a complex system collectively. Using concepts and tools such as emergence, connection topology, measures of complexity, attractors and feedback loops which suggest intervention points, we are able to construct Aotearoa's information landscape as an ecosystem, and develop

deeper understandings of networked harm, long-tail effects, emerging themes and trends, and clustered actors and agents. A complex systems approach, which originates in the natural science has been justifiably critiqued since "much of what matters in systems involving humans is left out" with little ability to understand the role of power, values, culture, politics or choice (Gandar, 2015). We have developed a mixed methods approach which enables exploration of those critical human systems through close, distant, and scalable reading.

A.6 Mixed Methods

Combining qualitative and quantitative research methods within a feminist, postcolonial framework, we build rich and diverse datasets to be explored both computationally and qualitatively. Using social network analysis combined with text-based analyses such as narrative analysis, discourse analysis, and reflexive thematic analysis, we map an ecosystem of disinformation, dangerous speech, harassment and abuse, and hateful and violent expression which is deeply grounded in the lived experiences of those most frequently targeted in Aotearoa New Zealand. Narrative analysis is a qualitative research method which enables interpretation of stories told within the context of research. It often draws on diverse texts, both collected (ie interviews) and found (ie social media posts). Understanding narrative, and how its role or performance has changed over time, is essential as a tool of both knowledge gathering and dissemination, while also enabling the voices of those targeted to be centered (Nadar, 2014). Discourse analysis is a qualitative research method which develops interpretations of language within social contexts. Sociopolitical discourse analysis "focuses on the social construction of discursive practices that maintain the social context", and specifically, critical discourse analysis (CDA) focuses on "the role of discourse in the (re)production and challenge of dominance (Salkind, 2010). Thematic analysis is a set of qualitative research methodologies which focus on identifying patterned meaning across a dataset (Braun, 2013). Reflexive thematic analysis is theoretically flexible, and grounded in researcher subjectivity. These mixed methods approaches are iterative and recursive – themes identified during narrative, discourse, and thematic analysis inform further areas of computational and qualitative exploration.

The rich and text-heavy dataset approach we use here combines the ability of "distant reading" to uncover or expose previously unseen patterns in a social network, with "reading against the grain" and close reading of "double-voiced texts" within archival and published sources (Moretti, 2013), (Bartholemae, 1993), (Gilbert, 1979). These mixed methods are predicted to reveal more and different relationships between key people, places, spaces, and artefacts or content, over time, than either merely distant reading or close reading can. Drawing on Martin Mueller's notion of "scalable reading", which offers "new and powerful ways of shuttling between 'text' and 'context'", we see that the deeply contextual theoretical approaches of feminist criticism and postcolonial theory enrich scalability (Mueller, 2012).

A.7 Study Definitions

We use the following definitions from Berentson-Shaw and Elliot (2020). (Note these definitions differ from those used in the body of the report, that are given in Section 2.2.4.)

Misinformation: "false information that people didn't create with the intent to hurt others"

Disinformation: "false information created with the intention of harming a person, group, or organization, or even an company"

Malinformation: "true information used with ill intent" (Shaw, 2020)

We draw from Dentith's (2018) work for a simple definition of conspiracy theories, defining them as purported explanations which cite a conspiracy at the salient cause of some event or phenomenon.

We use the definition of dangerous speech described by Susan Benesch at the Dangerous Speech Project:

Dangerous speech is any form of expression (speech, text or images) that can increase the risk that its audience will condone or participate in violence against members of another group.

We also draw on the Dangerous Speech Framework, and the hallmarks of dangerous speech.[5]

A.8 Literature Review: Deliberative democracy in countering online disinformation

The recent surge of online mis/disinformation[6] following the 2016 U.S. election, Brexit, and more recently, the COVID-19 pandemic is embedded in diminishing trust in Western democratic institutions (see Bennett and Livingstone, 2020; Freelon and Wells, 2020). Indeed, our research to date regarding online COVID-19 misinformation (Soar et al., 2020) and vaccine hesitancy have shown that there is lack of trust in government agencies in a vast range of government institutions including public health and medicine, particularly among marginalised communities in society. The spread of misinformation stokes issues of epistemic cynicism and uncertainty (i.e., citizens having difficulties in differentiating facts from mal-information and determining the reliability of the sources) (Chambers, 2021; McKay and Tenove, 2021). It also contributes to widening the existing socio-political cleavages, and weakening of social cohesion (Chambers, 2021; McKay and Tenove, 2021; Monsees, 2021; Tenove, 2020).

Within this context, the practices of public deliberation have been suggested as a tool for countering and mitigating impacts of online misinformation (Chambers, 2021; McKay and Tenove, 2021; Wikforss, 2020). Public deliberation often plays an important role in restoring citizens' trust and sense of belonging in political institutions as well as promoting social cohesion (Gastil, Black, and Moscovitz, 2008; Monsees, 2021). Within this framework, democracy is not reduced to election and technocratic decision-making in concert with the state bureaucracy, rather the emphasis is given to interactions between different spaces of communication in a political system (McKay and Tenove, 2021). Deliberative democracy is "a normative theory of democratic legitimacy based on the idea that those affected by a collective decision have the right, opportunity, and capacity to participate in consequential deliberation about the content of decisions" (Ercan, Hendriks, and Dryzek, 2019, p. 23). In deliberative systems, "citizens come together in face-to-face designed settings with good information, trained moderators, and procedural norms that promote participant equality in the deliberative and decision-making process" (Chambers, 2021, p. 152).

In the era of an information disorder[7], post-truth predicaments[8], diminishing public trust in democratic institutions and increased polarization, a system of deliberation might be imperative for sustainable democracy (Chambers, 2021; McKay and Tenove, 2021; Monsees, 2021). In this regard, we must recall the "truth-tracking potential" of deliberation (Habermas, 2006, p. 114). Deliberation as a structural and procedural matter has a logic that pushes toward the truth (Habermas, 2006). The core features of deliberation such as free debate, equal status of citizen, a critical public sphere, the circulation of information, and pluralism are the structural conditions that push toward better answers to questions and better solutions to problems through enhancing the legitimacy of decision making processes (Chambers, 2021; Habermas, 2006).

Designing spaces and moments for improved listening and reflection is the key feature of deliberation (Ercan et al., 2019). A deliberative system consists of differentiated yet linked communicative spaces that might range from highly structured forums (such as legislatures) to loose informal social gatherings and public interactions (Ercan et al., 2019). Deliberative democratic processes come in a variety of forms and sizes including citizen's juries, citizen panels and assemblies, deliberative forums and polls and consensus conferences or tribunals (Dryzek et al., 2019; Newton, 2018). However, all these forms share some key elements. A deliberative initiative involves diverse participants (randomly or non-randomly selected), facilitated dialogue, and an emphasis on norms of civility, careful institutional design and participant diversity (Dryzek et al., 2019). People involved

[5]Dangerous Speech: A Practical Guide. The Dangerous Speech Project: 19 April 2021 https://dangerousspeech.org/guide/

[6]Misinformation refers to all forms of false and inaccurate information that is not created with an intent to harm others. In comparison, "Disinformation entails the purposeful dissemination of falsehoods towards a greater dubious agenda and the chaotic fracturing of a society" (Silva et al., 2020).

[7]Wardle and Derakhshan (2017) used the term "information disorder" to refer to mis/dis and mal information.

[8]Where appeals to emotion are dominant and factual rebuttals or fact checks are ignored on the basis that they are mere assertions (see, Suiter, 2016).

in designing and facilitating deliberative initiatives need to be conscious of the framing – the ways in which information and messages are defined and presented to the participants (Blue and Dale, 2016). For a robust deliberation, it is important to ensure that framing is not leading the participants toward a certain outcome by making selective information visible. For this reason, it is particularly useful to involve people from diverse backgrounds. Participants should have an opportunity to "explore and examine alternative frames of an issue, and to confront each other with rival world-views, competing ideals, and conflicting political commitments" to avoid any premature agreements or false consensus (Blue and Dale, 2016, p. 3). Moreover, the bureaucracy that would otherwise oversee this policy area be committed to consider the recommendation made by the participants when making decisions. The deliberation must be well-resourced to allow the participants enough time to engage with the question: they are given enough time and access to relevant knowledge and experts (often of their choice); they're paid for the hours spent; food and adequate space is provided. And finally, then, they are expected to find a shared ground: perhaps not the solution they are the most passionate about, but something they can all "live with" (Buklijas, 2021).

Despite the scepticism about the deliberative system as a utopia, there are several examples of high-quality public deliberations that have successfully been used in different areas of decision-making (see, Curato, Dryzek, Ercan, Hendriks, and Niemeyer, 2017; Dryzek et al., 2019; Ercan et al., 2019; Gastil et al., 2008). Public deliberation has considerable appeal in making policy decisions that are complex, sensitive, and polarised (Gastil et al., 2008). Issues such as the criminal justice system (Dzur and Mirchandani, 2007), public health (Abelson, Blacksher, Li, Boesveld, and Goold, 2013; McWhirter et al., 2014), climate change (Blue, 2015; Blue and Dale, 2016; Sandover, Moseley, and Devine-Wright, 2021), and disinformation (Chambers, 2021; Monsees, 2021) are subjected to public rational-critical debates. Deliberation has also been successfully used in areas of social reform. For instance, the Irish Constitutional Convention and Citizen's Assembly convened to deliberative same-sex marriage and abortion, both of which ultimately led to legalisation of such practices. Such processes are also used to legitimise government. In Mongolia, any constitutional amendment now has to be preceded by a deliberative poll involving several hundred ordinary citizens (Dryzek et al., 2019). The world's largest deliberative institution is state-mandated village assemblies (gram sabha) in India (Dryzek et al., 2019).

A.9 A situated and contextual approach of deliberation for Aotearoa

Given that marginalised people are more likely to be susceptible to mis/disinformation (see, Hannah, 2020; Ngata, 2020), Aotearoa should adopt a situated and contextual approach. In other words, instead of importing the Western models of engagement, the deliberative practice in Aotearoa should consider an approach that "would combine the best elements of internationally tested deliberative democracy with Tiriti o Waitangi obligations and a commitment to biculturalism, while also sensitive to growing multiculturalism" (Buklijas, 2020, para. 7). This would mean recruiting a group of citizens that represents demographically the relevant population (of a neighbourhood, town, city, or whole country) while ensuring that the question asked of the group is framed using te ao Māori values and Māori knowledge-holders strongly represented among experts (Buklijas, 2020, 2021).

Public deliberation is not without its problems, however. On matters related to resource use allocation in New Zealand, the constituents who tend to take part are often the most privileged, with the money, time, education and confidence to participate (Buklijas, 2021). For instance, Newton (2018) showed that the demographic composition of submitters to the Auckland Plan 2050 plan did not match the demographic composition of the urban population – the "loudest voices" were older, wealthier, and whiter - than our city is today, let alone representative of a 2050 Tāmaki Makaurau. This implies that deliberation may result in reinforcing societal inequality if it is not designed to encourage the inclusion of the marginalised voices (see, Holdo and Öhrn Sagrelius, 2020; Siu, 2017).

Further to deliberative processes at the local level, while the deployment of deliberative processes to counteract online disinformation is scarce, there are examples where deliberative processes have been (mis)used in areas where disinformation was rife. Analysis of such cases provide a clue about the merits of deploying deliberative processes to counteract disinformation. In the context of public health in Aotearoa New Zealand, decisions by local government on the continuation of community water fluoridation have at times drawn on tribunal processes which have led to outcomes that are at odds with majority thought (Sivaneswaran et al, 2010;

Wyman et al, 2015). Weak majority thought is when there is a comfortable majority support on a particular issue (e.g., fluoridation of the drinking water) but that majority lack an intensity of opinion on such an issue (Campbell et al., 1980). In 2011, New Plymouth District Council voted to discontinue community water fluoridation following a one-day tribunal and in 2013 Hamilton also followed. The tribunal outcome in Hamilton was dramatic with Hamilton City Council voting 7-1 to discontinue community water fluoridation despite two previous referendums suggesting there was still widespread public support for the practice (Watson and Mace, 2014).

For deliberative processes in Aotearoa New Zealand to be deployed successfully, more will likely be required to overcome the consequences of weak majority thought. This is especially true in matters of public health where support or opposition for a particular public health measure seems divided into two camps claiming a superior knowledge of the science (Winstanley, 2005). That is, deliberative processes will need robust controls. These might include having an independent arbiter: a scientifically literate professional or professionals who already have broad legitimacy in the community and can and comment on the validity of claims by participants from an objective standpoint (see Oldfield, 2016).

A.10 Deliberative, iterative, ongoing community consultation

> We have a history of people putting Māori under a microscope in the same way a scientist looks at an insect. The ones doing the looking are giving themselves the power to define.
> - Merata Mita (On New Zealand Listener, 1989).

> I would like to insist on the embodied nature of all vision and so reclaim the sensory system that has been used to signify a leap out of the unmarked body and into a conquering gaze from nowhere. This is the gaze that mythically inscribes all the marked bodies, that makes the un-marked category claim the power to see and not be seen, to represent while escaping representation. This gaze signifies the un-marked positions of Man and White ...
> - Donna J. Haraway (1988)

Drawing from ongoing work to develop deliberative processes for New Zealanders, particularly those who are most targeted, scapegoated, and/or harmed by disinformation, dangerous speech, online harassment and abuse, and hateful or violet expression, in May 2021 we began piloting an approach to enable genuine, long-term and relationship-focused collective engagement between communities, government, civil society organizations and industry groups, prior to opening that collective engagement approach to include social media companies.

Community consultation often takes place at the end of a project, or at the beginning, to get approval, with no ongoing relationship between the entities delivering work for and on behalf on communities. The social media governance project embeds community consultation as an ongoing and normative relationship-based process which benefits all parties – in this case, social media companies, governments, and communities. We also focused not on representation, but on intersectionality and diversity of experience. In Aotearoa New Zealand, consultation between government and communities on key issues sometimes sees official civil society groups become spokespeople for a community, without necessarily being mandated by all members of that community. A broad, intersectional, diverse, accessible and inclusive multi-stakeholder forum methodology, which we have instituted, means that minority viewpoints within underserved communities are expressed, the burden of engagement (time, emotional labour, energy, costs) are more widely distributed, and the balance of power lies with the communities who have generously provided their time and knowledge.

Consensus-building requires localised, situated solutions to universally experienced harms and issues, and well-designed spaces and processes within which this can take place. Focusing on building a co-created approach, in which policy-makers and civil society/industry organizations are primed to listen more than they speak, an initial hui (meeting) was held in June 2021, with broad and diverse community representation from Māori, Pasifika, faith-based, ethnic, refugee and migrant communities, civil society and industry organizations, the social media governance community consulation/Disinformation Project team, who hosted, and representatives of the New Zealand Government. Participants' travel, accommodation and other costs were covered.

Exploring the complex interrelations between social media as a tool for connection, information, and knowledge flows on the one hand, while also being spaces and places within which harmful usage and misuse are enabling the production and amplification of hate and division on the other, we collectively discussed a design and development process.

This co-creation approach meant that in this initial hui (meeting) it was quickly identified that many of those present expressed that over the last three-four years they or other members of their community had contributed to a number of local and national government consultations or made submission on proposed legislative changes for new regulations or laws related to this general area of online harm, hate speech, and regulation of media. Many also expressed that they had also participated in workshops or meetings with civil society and industry organizations repeatedly over the same time period, including direct engagement with social media companies or other international working groups (The Christchurch Call, Global Internet Forum to Counter Terrorism). Many participants expressed how previous consultations, submissions, or participation in external workshops or working groups expected unpaid labour, unequal knowledge exchange, and were not ongoing. They also expressed how each experience of engagement started from the beginning, with little or no capture of previous consultations or submissions. Their grounded, situated knowledges were being sought, but not then used to develop better processes.

As such, we immediately identified that co-creation as first principle is essential for the development of a set of inclusive, responsive and iterative, survivor-led and community centred protocols for classifying harm experienced on social media platforms. We undertook then to include in our project a new strand – the gathering of as much data about related community consultations, submissions, workshops, and working groups as is publicly available, and a thematic analysis of the text we gather. We are now collating all published or submitted material provided by community groups to the New Zealand government over the period of the last 4 years in relation to hate speech, media regulation and social media regulation, social cohesion, the Internet, and the Christchurch mosque terrorist attack. The purpose of this is threefold: 1) to identify which groups have been consulted with in order to identify gaps or silences, and to mitigate the effects of over-consultation; 2) to undergo a thematic analysis of the content of the submissions to inform government and civil society/industry organization processes; 3) and, most importantly for the social media governance workstream, to establish a set of thematic baselines for the next face-to-face community consultations.

Aotearoa New Zealand has been experiencing a Delta outbreak, centred in Tāmaki Makaurau Auckland, our largest city, since the middle of August, and the key priority for communities during this period, and ongoing, is increasing vaccination rates and mitigating the effects of Covid-19 mis- and disinformation, most of which is produced and amplified via social media platforms, on increasing vaccine hesitancy or vaccine resistance. Closely working with communities, the social media governance community consultation team have been monitoring and analysing publicly available social media mis- and disinformation related to Covid-19 and vaccination in particular, to inform communities who are then empowered to develop community-led responses to targeted disinformation. We have also worked closely with Netsafe and other government and non-governmental agencies to, where appropriate, report mis- and disinformation to social media platforms for review against existing community guidelines and Covid-19 policies. Given the immediate and ongoing nature of this aspect of the project, with personal security risks to community consultation participants and researchers engaged with this work, detail of this active research will be reported to GPAI in 2022.

Further small group online 'zui' (zoom meetings) have been held over August-September 2021, to hone the processes for a next, larger scale meeting, and to ensure ongoing connection maintained. Another round of larger scale online meetings is scheduled for late October, given the ongoing inability to meet in person, which is the preference for trust establishment and consensus-building.

A.11 Preliminary Recommendations

The approach undertaken in the social media governance community consultation project seeks to be grounded in the Universal Declaration of Human Rights, but also within foundational legal, civil, and social frameworks that are particular to Aotearoa New Zealand, critically Te Tiriti O Waitangi.

Linda Tuhiwai Smith's set of questions (Smith, 2012) that need to be asked in a cross-cultural research context must form the basis for the development of the process:

- Who defined the research problem
- For whom is this study worthy and relevant? Who says so?
- What knowledge will the community gain from this study?
- What knowledge will the researcher gain from this study?
- What are some likely positive outcomes from this research?
- What are some possible negative outcomes?
- How can the negative outcomes be eliminated?
- To whom is the researcher accountable?
- What processes are in place to support the research, the researched and the researcher?

In establishing locally appropriate norms within an international human rights-focused framework, we aimed to contribute to principles for growing international consensus, enabling social democracies to act to build social cohesion in ways that are situated, locally grounded, respectful of rights and responsibilities, and operate within existing agreed principles of international law and regulation.

1. Well-designed deliberative democratic processes, with a focus on co-creation, can work to establish relationships and trust between communities, government, and civil society/industry organizations, and can be utilized in social democracies in order to start grounded, situated protocols for working with social media companies.
2. Researchers, social media companies, governments, and civil society/industry organizations need to be committed to deliberative democratic processes for community consultation in an iterative and ongoing manner – this is not once-off, nor is it set in stone.
3. Over- and under-consultation with communities is likely; researchers, governments, and civil society/industry organizations should collate and analyse publicly available published or submitted data as part of the establishment phase for deliberative democratic processes with communities in their respective jurisdictions. This will both provide a far more detailed baseline, and enable identification of gaps or silences in previous processes.
4. Social media companies, researchers, governments, and civil society/industry organizations must be prepared to listen more than they speak, and to acknowledge that communities are the experts in their own lives and lived experiences of disinformation, dangerous speech, abuse and harassment, and hateful and violent expression.
5. It is these situated lived experiences, which will differ for individuals, communities, and countries, which will better inform social media spaces as to what constitutes disinformation, dangerous speech, abuse and harassment, and hateful or violent expression for this community, in this place, at this time.
6. participate. The deliberation must be well-resourced to allow the participants enough time to engage with the question: they are given enough time and access to relevant knowledge and experts (often of their choice); they're paid for the hours spent; food and adequate space is provided.

B A review of current platform methods for guiding 'user journeys'

The concern driving our project is that social media recommender systems may be causing various types of harmful content to proliferate on platforms. Of course, social media companies are aware of the problem of harmful content on their platforms, and take many steps to address the problem. In this annex, we will review what we know about what social media companies have done, and are doing, to address the problem.

Broadly speaking, there are two possible ways companies can reduce the amount of harmful content on platforms. One way is to remove harmful content. We will discuss what we know about projects for removing harmful content in Section B.1. Another way is to reduce the dissemination of harmful content, or increase the dissemination of content that combats harmful content. We review what we know about company initiatives in this area in Section B.2, focusing on methods for 'downranking' certain categories of 'borderline content' that falls short of the criterion for removal, and for upranking 'authoritative content'. We applaud all these initiatives—but we argue that they certainly don't obviate the need for the kind of fact-finding exercise we proposed in Chapter 5. We argue that methods for removing harmful content are far from sufficiently accurate, so manipulations of the recommender system are likely to play an important role in keeping harmful content off platforms. We argue that methods for down/upranking content may well be informed by the kind of study we propose—but that information about the motivating studies should be surfaced in the public domain.

B.1 Companies' initiatives for removing harmful content

We know very little about the technical methods social media platforms use to classify content. What little we know was summarised in Section 2.6.1.2. We think this lack of transparency is increasingly problematic, given the present-day reach of these platforms. Indeed, one of the aims of our community consultation project (Annex A) is to explore ways of making some aspects of the process of classifying harmful content more transparent. However, companies do make certain public statements about the categories of harmful content they remove, and how well their current methods perform. In this section we will review these statements.[1]

B.1.1 Facebook/Instagram

Facebook and Instagram's Community Standards policy[2] bans several categories of harmful content, relating to 'Violent and criminal behaviour' (including 'Violence and incitement', 'Dangerous individuals and organisations'), 'Safety' (including 'Suicide and self-injury', 'Child sexual exploitation, abuse and nudity'), 'Objectionable content' (including 'Hate speech', 'Violent and graphic content'), and 'Authenticity' (including 'False news').

The main metric Facebook uses to publicly report the success of its classifiers is the 'proactive detection rate', which is the percentage of violating content found by its classifiers before it is reported. Facebook currently reports a proactive detection rate of over 90%. This number varies over time and across presentations: for instance, a November 2020 news story on 'Combating Hate Speech'[3] gives a figure of 95%. But it would be much more meaningful if Facebook reported the performance of its classifiers on a corpus which has been annotated by hand: this is the gold-standard test for evaluating AI classifiers. It would also be useful to know how often Facebook acts on the violating content it automatically identifies.

[1] A few more details about the performance of classifiers can be found in the EU's recent summary of information on this topic provided to them by companies (European Commission, 2021).

[2] https://transparency.fb.com/en-gb/policies/community-standards

[3] https://about.fb.com/news/2020/11/measuring-progress-combating-hate-speech/

As noted in Section 2.6.1.2, the whistleblower Frances Haugen reports internal studies indicating that Facebook 'may action as little as 3–5% of hate and about 6-tenths of 1% of V & I [violence and incitement]'. The difference between these figures and the figure given publicly may indicate that the proactive detection rate metric doesn't tell us everything we need to know. But as yet we really don't know.

B.1.2 Twitter

Twitter's harmful content policy[4] uses the following categories to define harmful content: 'Safety' (including violence, terrorism and violent extremism, child sexual exploitation, abuse and harassment, hateful conduct, suicide or self-harm, sensitive media, including graphic violence and adult content), 'Privacy' (including 'private information'), 'Authenticity' (including civic integrity).

Like Facebook, Twitter's public discussion of content moderation often reports the rate at which content is identified and removed before being reported. For instance, the rate for abusive content was reported as 'more than 50%' in Twitter's 2019 Q3 shareholder letter (Twitter, 2019). But Twitter often reports absolute numbers—for instance, its 2021 report to the EU describes the number of accounts actioned and suspended, and the number of content items removed (European Commission, 2021). These statistics tell us far less about the actual performance of Twitter's classifiers.

Whatever its performance, there is interesting evidence that Twitter's content moderation is highly tunable. Andres and Slivko (2021) discovered that introduction of the German Network Enforcement Act (NetzDG)—which requires rapid removal of types of harmful content—was followed by a reduction of around half the median number of hateful Tweets per month in certain content categories within German Twitter. The authors note, however, that this reduction may not be entirely due to Twitter's proactive content removal: the introduction of NetzDG may have also increased the number of hateful Tweets reported by users, as well as discouraging users to post hateful Tweets in the first place.

B.1.3 YouTube

YouTube's harmful content policy[5] includes categories for 'Harmful or dangerous content' (including 'Instructions to harm' and 'Eating disorders'), 'Violent or graphic content' (including incitements to violence and scenes of violence), 'Violent criminal organisations' (including including content produced by, praising or recruiting for terrorist organisations) and 'Hate speech' (targeted at individuals, or at a variety of groups).

YouTube reports performance of its classifiers using a mixture of absolute numbers and variants of the proactive detection rate. For instance, on its Transparency Report page,[6] users can ask for the number of videos removed by 'automated flagging' for a given time period; in its recent report to the EU, it states that 'approximately 76% of the videos uploaded that were removed for violating our Hate Speech policy were taken down before they had 10 views' (EU, 2021).

Interesting reports surfaced recently about the balance struck by YouTube between human and automated content classifiers. A Wall Street Journal article (Barker and Murphy, 2020) reported that YouTube shifted toward significantly more use of AI moderation during the early parts of the COVID-19 pandemic, and this resulted in a doubling of videos being taken down in the second quarter of 2020. However the high false positive rate was deemed unacceptable, and the number of human moderators was increased, in order to improve the precision of moderation outcomes.

Since April 2021 Google has been publishing a statistic about YouTube that it developed: the 'violating view rate'. This statistic indicates the proportion of videos viewed within which the videos violate any of YouTube's rules. At the time the metric was introduced,[7] its value was reported to be below 0.2%. However one concern

[4] https://help.twitter.com/en/rules-and-policies/twitter-rules
[5] https://support.google.com/youtube/topic/2803176?hl=en&ref_topic=6151248
[6] https://transparencyreport.google.com/youtube-policy/removals?hl=en
[7] https://www.vox.com/recode/2021/4/6/22368809/youtube-violative-view-rate-content-moderation-guidelines

is that in absolute terms, this is still a lot of viewing activity, globally. Another concern is that the value is reached by sampling played views and sending that sample for content classification by in-house staff, which may lead to (potentially unintentional) bias.

B.1.4 Do companies' content removal protocols address concerns about recommender systems?

The main concern we articulate in this report is that social media *recommender systems* may have a role in spreading or amplifying harmful content of various kinds. One response companies might make is to point to their mechanisms for *removing* harmful content: the classifiers we have been discussing in the current section. If these classifiers were good enough, it could perhaps be argued that concerns about recommender systems are unfounded. Our main reason for reviewing what we know about the performance of these classifiers is to head off that argument.

On the one hand, companies don't publicly report the performance metrics that we would really need to assess this question. Absolute numbers don't tell us enough. Neither do proactive detection rate measures. Furthermore, reports from whistleblowers like Frances Haugen tell a very different story the accuracy of classifiers. Again, we just don't know enough in this area.

More importantly, we can always expect a fairly large measure of error in the performance of classifiers. If there are manipulations of recommender systems that reduce the extent to which harmful content is *circulated*, or *amplified* on a given platform, this strategy for reducing the amount of harmful content on the platform is one that should be given careful thought. Otherwise, companies find themselves playing 'whack-a-mole'—a metaphor that has circulated widely in the wake of Haugen's recent testimony before Congress.

Note that the fact-finding exercise we advocate in Chapter 5 addresses exactly the question of whether there are manipulations of the recommender algorithm that reduce the circulation of harmful content. Its aim is precisely to measure how variations in the recommender system affect users' relationship towards harmful content.

B.2 Downranking 'borderline content' and upranking 'authoritative content'

The second approach companies take towards harmful content is to downrank it. The process of downranking essentially tweaks the 'scores' of content items computed by the recommender system (as described in Section 1.1.1). Scores for other kinds of content can be tweaked in the other direction: often, 'authoritative content' is promoted, which implicitly downranks other content.

Again, we don't know how this tweaking is done. In particular, we don't know whether it is done by a postprocessor operating on the output of the recommender system, or by a recommender system whose architecture has been modified (for instance, with different input fields, or different optimisation objectives).

What companies describe about their changes to recommender systems relates almost entirely to *borderline content* (for downranking), and to *authoritative content* (for upranking). In this section we will review what is known about how companies define these categories of content, and about the effects of companies' manipulations to recommender system scores for these content categories.

B.2.1 YouTube's recommender system modifications

A YouTube blog post (Godrow, 2021) usefully traces the history of modifications to the YouTube recommender algorithm, to demote 'sensationalist' and 'borderline' content and promote 'authoritative news and information'.[8] Other earlier posts claim quantified improvements: for instance, YouTube (2019) claims a 70% drop in

-spam-hate-speech

[8] This post also includes manipulations that block 'racy or violent' videos, which count as content removal protocols.

watch time of 'borderline content and harmful misinformation' on the platform, after a raft of changes to the recommender algorithm. But this figure incorporates results from content removal protocols as well as tweaks to the recommender system.

Faddoul *et al.* (2020) conduct a useful check on YouTube's quantitative claims, for one specific category of problematic content, 'conspiracy content'. YouTube's claim to have lowered watch time of harmful content by 70% by 2019 is roughly validated. However, Faddoul *et al.* find a subsequent rebound since that date, with a much lower net reduction of 40% at the time of their 2020 study.

We know very little about how borderline content is defined. The definition probably identifies content that 'comes close' (in the judgement of those curating training sets) to one of the categories of harmful content that is straightforwardly banned. This raises various questions—but as we have already discussed in Section 2.2.3, there are well-argued cases for classifying harmful content on a continuous scale of severity (see again Aroyo *et al.*, 2019; Kritchenko and Nejadgholi, 2020). It might be useful for companies to make samples from their training set of borderline content—though of course the sample size should be small enough that it doesn't enable adversarial methods.

We know a lot more about YouTube's actions in promoting 'authoritative content', since these largely involve improved (sometimes formal) relationships with conventional news organisations. Again these actions are usefully summarised in one of YouTube's blog posts[9].

B.2.2 Facebook's recommender system modifications

On September 22, 2021, Facebook updated their Transparency Center website regarding the types of content that is demoted within each user's News Feed.[10] The types of content include spam-related categories (content from 'ad farms' and those seeking to monetise 'low-quality videos'), and a category of 'fact-checked misinformation'. It also includes several categories of content whose downranking 'fosters safer communities'. Among these latter categories, there is an interesting distinction between 'content borderline to our community standards' (which is identified as such by human annotators) and 'content likely to violate our community standards' (which is content *automatically classed* as violating standards *prior to confirmation* by human annotators). Facebook are erring on the side of caution in this second case—but to what extent we don't know.

Facebook has also promoted various categories of 'authoritative content' at various points in the past. During the Covid crisis, it tweaked the feeds of users who engaged with fake Covid news to include 'anti-misinformation messages' (see e.g. Robertson, 2020). In the aftermath of the US 2020 election, Facebook appeared to make a more systematic change to its recommender, in an attempt to stem the surge in misinformation about missing ballots and related issues (see e.g. Summers, 2020). According to Summers' article, Facebook 'increased the weighting of an internal value called NEQ (news ecosystem quality), which tries to quantify the value and rigor of an outlet's journalism'; this feels like a tweak to the recommender system's parameters, rather than just a postprocessing of its output scores. Interestingly, Facebook apparently 'rolled back' these changes before the end of 2020. (Frances Haugen refers to a similar rollback of recommender system parameters in her recent testimony.)

B.2.3 Twitter's recommender system modifications

Twitter notes in its Help Center page entitled 'Our range of enforcement options'[11] that downranking is a possibility, under the heading 'Limiting Tweet visibility'. In Twitter's hierarchy of processes, downranking a Tweet is a step up from labelling it as 'containing disputed or misleading information', and a step down either requiring its removal, or going further to hide the Tweet in question while waiting for the poster to remove it.

[9] https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce/
[10] https://transparency.fb.com/en-gb/features/approach-to-ranking/types-of-content-we-demote/
[11] https://help.twitter.com/en/rules-and-policies/enforcement-options

As already noted in Section 2.6.1.2, one case in which the above downranking may not be applied is when it is considered to be in the public interest for a Tweet to remain accessible, such as happened with some of Donald Trump's Tweets. But in such cases the Tweet will nonetheless be moved behind a notice from Twitter.

B.2.4 How do companies' current approaches to down/upranking bear on our proposed fact-finding exercise?

Companies' modifications to ranking processes probably happen for a variety of reasons: they may be responding to current circumstances, or trying out various policies. They may also happen through a variety of mechanisms, some involving adjustment to parameters, and some involving post-processing measures.

It may be that some modifications are informed by the kinds of study we advocate. For instance, Facebook's varying of the 'NEQ' parameter may well have been done with the knowledge of the effects this would produce. If this is the case, it would indicate that many of the analytic tools for our proposed fact-finding exercise may well already be in place—and we would imagine the same would be true of other companies. But again, this emphatically does not remove the need for the fact-finding exercise. The exercise we propose will not only *perform* studies about recommender system effects, but will also *communicate* these studies, and their results, to the public. Once the results of such studies are known, it may well be that certain principles of good practice or social responsibility can be defined about how to choose the right recommender system parameters for a given circumstance. We think the fact-finding exercise may surface findings that contribute usefully to policy discussions on this matter.

C Legal/policy issues for the proposed fact-finding exercise

Our survey of legal and policy issues for the proposed fact-finding exercise was conducted by Tom Barraclough and Curtis Barnes (Barraclough and Barnes, 2021). The report appears under separate cover: it can be found at http://www.brainbox.institute/GPAIproposal.

References

Abelson, J., A., B. E., K., L. K., Boesveld, S. E., & Goold, S. D. (2013). Public deliberation in health policy and bioethics: mapping an emerging, interdisciplinary field. *Journal of Public Deliberation*, *9*(1).

Aberson, C. L., Healy, M., & Romero, V. (2000). Ingroup bias and self-esteem: A meta-analysis. *Personality and Social Psychology Review*, *4*(2), 157–173.

Alava, S., Frau-Meigs, D., & Hassan, G. (2017). *Youth and violent extremism on social media: Mapping the research.* UNESCO report.

Allcott, H., Braghieri, L., Eichmeyer, S., & Gentzkow, M. (2020). The welfare effects of social media. *American Economic Review*, *110*(3), 629–76.

Allcott, H., Gentzkow, M., & Yu, C. (2019). Trends in the diffusion of misinformation on social media. *Research & Politics*, *6*(2), 2053168019848554.

Allen, C. E. (2007). *Threat of Islamic radicalization to the homeland.* Written testimony of Charles E. Allen, Assistant Secretary of Intelligence and Analysis, Department of Homeland Security, before the US Senate Committee on Homeland Security and Governmental Affairs, p.4.

Aluru, S. S., Mathew, B., Saha, P., & Mukherjee, A. (2020). Deep learning models for multilingual hate speech detection. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD).*

Anastasio, N., Perliger, A., & Shortland, N. (2021). How emotional traits and practices lead to support in acts of political violence. *Studies in Conflict and Terrorism*, 1–21.

Anderson, P. (1972). More is Different. *Science*, *177*(4047:393-396).

Andres, R., & Slivko, O. (2021). The effect of hate speech regulation on German Twitter. *Available at SSRN 3896399*.

Arendt, F., Scherr, S., & Romer, D. (2019). Effects of exposure to self-harm on social media: Evidence from a two-wave panel study among young adults. *New Media & Society*, *21*(11-12), 2422–2442.

Aroyo, L., Dixon, L., Thain, N., Redfield, O., & Rosen, R. (2019). Crowdsourcing subjective tasks: the case study of understanding toxicity in online discussions. In *Companion Proceedings of the 2019 World Wide Web Conference* (pp. 1100–1105).

Asimovic, N., Nagler, J., Bonneau, R., & Tucker, J. A. (2021). Testing the effects of Facebook usage in an ethnically polarized setting. *Proceedings of the National Academy of Sciences*, *118*(25).

Backstrom, L., Huttenlocher, D., Kleinberg, J., & Lan, X. (2006). Group formation in large social networks: Membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 44–54).

Bahador, B. (2020). *Classifying and identifying the intensity of hate speech.* Insights from the Social Sciences. https://items.ssrc.org/disinformation-democracy-and-conflict-prevention/classifying-and-identifying-the-intensity-of-hate-speech/.

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132.

Banko, M., MacKeen, B., & Ray, L. (2020). A unified taxonomy of harmful content. In *Proceedings of the Fourth Workshop on Online Abuse and Harms* (pp. 125–137).

Barker, A., & Murphy, H. (2020). *YouTube reverts to human moderators in the fight against misinformation.* Financial Times.

Barraclough, T., & Barnes, C. (2021). *Legal and public policy considerations: proposal by Global Partnership on AI for collaborative study of TVEC and social media recommender systems.* Brainbox report, http://www.brainbox.institute/GPAIproposal.

Bartholemae, D., & Petrosky, A. (1993). Ways of reading. *St Martins Press, Boston*.

Basile, V., Bosco, C., Fersini, E., Debora, N., Patti, V., Pardo, F. M. R., ... others (2019). SemEval-2019 Task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *13th International Workshop on Semantic Evaluation* (pp. 54–63).

Baugut, P., & Neumann, K. (2020). Online propaganda use during Islamist radicalization. *Information, Communication & Society*, *23*(11), 1570–1592.
Benesch, S., Buerger, C., Glavinic, T., Manion, S., & Bateyko, D. (2021). *Dangerous speech: A practical guide.* The Dangerous Speech Project.
Benitez-Parejo, N., del Águila, M. R., & Pérez-Vicente, S. (2011). Survival analysis and Cox regression. *Allergologia et Immunopathologia*, *39*(6), 362–373.
Bennett, L. W., & Livingstone, S. (2020). A Brief History of the Disinformation Age. In L. Bennett (Ed.), *The disinformation age: politics, technology and disruptive communication in the United States.*
Berentson-Shaw, J., & Elliot, M. (2020). Misinformation and Covid-19: a briefing for media. *The Workshop, Wellington*.
Blom, J. N., & Hansen, K. R. (2015). Click bait: Forward-reference as lure in online news headlines. *Journal of Pragmatics*, *76*, 87–100.
Blue, G. (2015). Public deliberation with climate change: opening up or closing down policy options? *Review of European, Comparative & International Environmental Law*, *24*(2: 152-159).
Blue, G., & Dale, J. (2016). Framing and power in public deliberation with climate change: Critical reflections on the role of deliberative practitioners. *Journal of Public Deliberation*, *12*(1:2).
Boduszek, D., Debowska, A., Sharratt, K., McDermott, D., Sherretts, N., Willmott, D., ... Hyland, P. (2021). Pathways between types of crime and criminal social identity: A network approach. *Journal of Criminal Justice*, *72*, 101750.
Bontcheva, K., & Posetti, J. (Eds.). (2020). *Balancing act: Countering digital disinformation while respecting freedom of expression*. Broadband Commission research report on 'Freedom of Expression and Addressing Disinformation on the Internet'.
Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., ... Snelson, E. (2013). Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, *14*(11).
Boudry, M., Blancke, S., & Pigliucci, M. (2015). What makes weird beliefs thrive? The epidemiology of pseudoscience. *Philosophical Psychology*, *28*(8), 1177–1198.
Bouneffouf, D., Laroche, R., Urvoy, T., Féraud, R., & Allesiardo, R. (2014). Contextual bandit for active learning: Active Thompson sampling. In *International Conference on Neural Information Processing* (pp. 405–412).
Boxell, L., Gentzkow, M., & Shapiro, J. M. (2017). Greater Internet use is not associated with faster growth in political polarization among US demographic groups. *Proceedings of the National Academy of Sciences*, *114*(40), 10612–10617.
Boxell, L., Gentzkow, M., & Shapiro, J. M. (2020). *Cross-country trends in affective polarization* (Tech. Rep.). National Bureau of Economic Research.
Brady, W. J., Crockett, M. J., & Van Bavel, J. J. (2020). The MAD model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science*, *15*(4), 978–1010.
Brady, W. J., McLoughlin, K., Doan, T. N., & Crockett, M. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances*, *7*.
Brady, W. J., & Van Bavel, J. J. (2021a). *Estimating the effect size of moral contagion in online networks: A pre-registered replication and meta-analysis.* OSF Preprints.
Brady, W. J., & Van Bavel, J. J. (2021b). Social identity shapes antecedents and functional outcomes of moral emotion expression in online networks.
Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318.
Brashier, N. M., & Schacter, D. L. (2020). Aging in an era of fake news. *Current directions in psychological science*, *29*(3), 316–323.
Braun, V., & Clarke, V. (2013). Successful qualitative research: a practical guide for beginners. *Sage*.
Brost, B., Cox, I. J., Seldin, Y., & Lioma, C. (2016). An improved multileaving algorithm for online ranker evaluation. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 745–748).
Bruns, A. (2019). Filter Bubble. *Internet Policy Review*, *8*(4).
Buklijas, T. (2020). Complex conversations: A deliberative democratic model for New Zealand. Retrieved

from https://informedfutures.org/complex-conversations/.
Buklijas, T. (2021). Using Deliberative Democracy to explore the future of Auckland's water supply. Retrieved from https://www.greaterauckland.org.nz/2021/10/06/using-deliberative-democracy-to-explore-the-future-of-aucklands-water-supply/.
Carlyle, K. E., Guidry, J. P., Williams, K., Tabaac, A., & Perrin, P. B. (2018). Suicide conversations on instagram™: contagion or caring? *Journal of Communication in Healthcare*, *11*(1), 12–18.
Chakraborti, T., Patra, A., & Noble, J. (2020). Contrastive fairness in machine learning. *IEEE Letters of the Computer Society*, *3*(2: 38-41).
Chambers, S. (2021). Truth, deliberative democracy, and the virtues of accuracy: is fake news destroying the public sphere? *Political Studies*, *69*(1:147-163).
Chang, P. F., & Bazarova, N. N. (2016). Managing stigma: Disclosure-response communication patterns in pro-anorexic websites. *Health Communication*, *31*(2), 217–229.
Chapelle, O., Joachims, T., Radlinski, F., & Yue, Y. (2012). Large-scale validation and analysis of interleaved search evaluation. *ACM Transactions on Information Systems (TOIS)*, *30*(1), 1–41.
Cho, J., Ahmed, S., Hilbert, M., Liu, B., & Luu, J. (2020). Do search algorithms endanger democracy? An experimental investigation of algorithm effects on political polarization. *Journal of Broadcasting & Electronic Media*, *64*(2), 150–172.
Comerford, M., Guhl, J., & Miller, C. (2021). Understanding the New Zealand online extremist ecosystem. *Institute for Strategic Dialogue, London*.
Conger, K., & Isaac, M. (2021). *Inside Twitter's Decision to Cut Off Trump.* New York Times.
Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL* (pp. 8440–8451).
Covington, P., Adams, J., & Sargin, E. (2016). Deep neural networks for YouTube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems* (pp. 191–198).
Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A Black Feminist critique of antidiscrimination doctrine, Feminist theory, and antiracist politics. *University of Chicago Legal*, *1*(139).
Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, *1*(11), 769–771.
Curato, N., Dryzek, J. S., Ercan, S. A., Hendriks, C. M., & Niemeyer, S. (2017). Twelve key findings in Deliberative Democracy research. *Daedalus*, *146*(3:28-38).
Deffuant, G., Neau, D., Amblard, F., & Weisbuch, G. (2000). Mixing beliefs among interacting agents. *Adv. Complex Syst.*, *3*, 87-98.
Del Vicario, M., Vivaldo, G., Bessi, A., Zollo, F., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2016). Echo chambers: Emotional contagion and group polarization on Facebook. *Scientific Reports*, *6*(1), 1–12.
de Vreese, C., Bastos, M., Esser, F., Giglietto, F., Lecleher, S., Pfetsch, B., ... Persily, N. (2019). *Public statement from the Co-Chairs and European Advisory Committee of Social Science One.* https://socialscience.one/blog/public-statement-european-advisory-committee-social-science-one.
Dryzek, J. S., Bächtiger, A., Chambers, S., Cohen, J., Druckman, J. N., Felicetti, A., & ...Gutmann, A. (2019). The crisis of democracy and the science of deliberation. *Science*, *363*(6432:1144-1146).
Dzur, A. W., & Mirchandani, R. (2007). Punishment and democracy: The role of public deliberation. *Punishment & Society*, *9*(2:151-175).
D'Ignazio, C., & Klein, L. (2020). Data Feminism. *MIT Press*.
Ercan, S. A., Hendriks, C. M., & Dryzek, J. S. (2019). Public deliberation in an era of communicative plenty. *Policy & Politics*, *47*(1:19-36).
European Commission. (2020). *Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC.*
European Commission. (2021). *Information provided by the IT companies about measures taken to counter hate speech, including their actions to automatically detect content.* EU Directorate-General for Justice and Consumers.
Evans, D. (2003). *Placebo: The belief effect*. Harper Collins London.
Evans, G., & King, G. (2020). Statistically valid inferences from differentially private data releases, with application to the Facebook URLs dataset. *Political Analysis. URL: GaryKing. org/dpd*.
Everett, J. A., Faber, N. S., & Crockett, M. (2015). Preferences and beliefs in ingroup favoritism. *Frontiers in Behavioral Neuroscience*, *9*, 15.
Faddoul, M., Chaslot, G., & Farid, H. (2020). A longitudinal analysis of YouTube's promotion of conspiracy

videos. *arXiv preprint arXiv:2003.03318*.
Finkel, E. J., Bail, C. A., Cikara, M., Ditto, P. H., Iyengar, S., Klar, S., … others (2020). Political sectarianism in America. *Science*, *370*(6516), 533–536.
Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, *80*(S1), 298–320.
Freelon, D., & Wells, C. (2020). Disinformation as political communication. *Political Communication*, *37*(2:145-156).
Frissen, T. (2021). Internet, the great radicalizer? Exploring relationships between seeking for online extremist materials and cognitive radicalization in young adults. *Computers in Human Behavior*, *114*, 106549.
Gandar, P. (2015). Book review: Handbook on complexity and public policy. *New Zealand Science Review*, *72*(4:10-106).
Ganesh, B. (2020). Evaluating the promise of formal counter-narratives. In B. Ganesh & J. Bright (Eds.), *Extreme digital speech: Contexts, responses, and solutions* (pp. 89–98). University of Groningen.
Gastil, J., Black, L., & Moscovitz, K. (2008). Ideology, attitude change, and deliberation in small face-to-face groups. *Political Communication*, *25*(1:23-46).
Gavaghan, C., Knott, A., Maclaurin, J., Zerilli, J., & Liddicoat, J. (2020). *Government use of Artificial Intelligence in New Zealand*. Wellington, New Zealand: New Zealand Law Foundation.
Gendron, A. (2017). The call to Jihad: Charismatic preachers and the Internet. *Studies in Conflict & Terrorism*, *40*(1), 44–61.
Geschke, D., Lorenz, J., & Holtz, P. (2019). The triple-filter bubble: Using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers. *The British Journal of Social Psychology*, *58*, 129 - 149.
GIFCT. (2021a). *Broadening the GIFCT Hash-Sharing Database Taxonomy: An Assessment and Recommended Next Steps.* Global Internet Forum to Counter Terrorism report.
GIFCT. (2021b). *Progress Continues for the Christchurch Call to Action.* https://gifct.org/2021/05/14/christchurch-call-to-action-second-anniversary/.
GIFCT. (2021c). *Report of the Content-Sharing Algorithms, Processes, and Positive Interventions Working Group Part 1: Content-Sharing Algorithms & Processes.* Global Internet Forum to Counter Terrorism.
Gilbert, S., & Gubar, S. (1979). The Madwoman in the Attic: The Woman Writer and the Nineteenth Century Literary Imagination. *Yale University Press*.
GLAAD. (2021). *GLAAD Social Media Safety Index.* Available from glaad.org.
Global Witness. (2021). *Algorithm of harm: Facebook amplified Myanmar military propaganda following coup.* Global Witness report.
Godrow, C. (2021). *On YouTube's recommendation system.* https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/.
Goffin, R. D., & Olson, J. M. (2011). Is it all relative? Comparative judgments and the possible improvement of self-ratings and ratings of others. *Perspectives on Psychological Science*, *6*(1), 48–60.
Gomez-Uribe, C. A., & Hunt, N. (2015). The Netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)*, *6*(4), 1–19.
Griffioen, N., van Rooij, M., Lichtwarck-Aschoff, A., & Granic, I. (2020). Toward improved methods in social media research. *Technology, Mind, and Behavior*, *1*(1).
Guess, A. M., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour*, *4*(5), 472–480.
Haas, E. J., Angulo, F. J., McLaughlin, J. M., Anis, E., Singer, S. R., Khan, F., … others (2021). Impact and effectiveness of mRNA BNT162b2 vaccine against SARS-CoV-2 infections and COVID-19 cases, hospitalisations, and deaths following a nationwide vaccination campaign in Israel: an observational study using national surveillance data. *The Lancet*, *397*(10287), 1819–1829.
Habermas, J. (2006). Political communication in media society: Does democracy still enjoy an epistemic dimension? The impact of normative theory on empirical research. *Communication theory*, *16*(4:411-426).
Hagey, K., & Horwitz, J. (2021). *Facebook tried to make its platform a healthier place. It got angrier instead.* Wall Street Journal.
Halpern, J., & Pearl, J. (2005). Causes and explanations: A structural-model approach. *The British Journal for the Philosophy of Science*, *56*(4: 843-911).
Hannah, K. (2020). Counting and countering the infodemic: A deep dive into Covid-19 disinformation. *The*

Spinoff.
Hao, K. (2021). *How Facebook got addicted to spreading misinformation.* MIT Technology Review, March 11.
Haraway, D. (1988). Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, *14*(3).
Haselton, M. G., Nettle, D., & Murray, D. R. (2015). The evolution of cognitive bias. *Handbook of Evolutionary Psychology*, 1–20.
Hmielowski, J. D., Beam, M. A., & Hutchens, M. J. (2016). Structural changes in media and attitude polarization: Examining the contributions of TV news before and after the Telecommunications Act of 1996. *International Journal of Public Opinion Research*, *28*(2), 153–172.
Hoffman, B., Ware, J., & Shapiro, E. (2020). Assessing the threat of Incel violence. *Studies in Conflict & Terrorism*, *43*(7), 565–587.
Holdo, M., & Öhrn Sagrelius, L. (2020). Why inequalities persist in public deliberation: Five mechanisms of marginalization. *Political Studies*, *68*(3:634-652).
Horgan, J. (2011). *Remarks at START symposium, 'Lessons learned since the terrorist attacks of September 11, 2001'.* Washington DC, 1 Sept. 2011, http://www.c-spanvideo.org/program/TenYearA.
Hosseinmardi, H., Ghasemian, A., Clauset, A., Rothschild, D. M., Mobius, M., & Watts, D. J. (2020). Evaluating the scale, growth, and origins of right-wing echo chambers on YouTube. *arXiv preprint arXiv:2011.12843*.
Hunter, J. A., Scarf, D., Trent, J., Hayhurst, J., Yong, M. H., Chan, J., … Stringer, M. (2019). Perceived control and intergroup discrimination. *Current Research in Social Psychology*.
Ilic, A., & Kabiljo, M. (2015). *Recommending items to more than a billion people.* Facebook Engineering blog post.
Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, *22*, 129–146.
Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems.*, *20*(4), 422–446.
Javed, J., & Miller, B. A. P. (2019). *The dangers of false news: How sensational content and outgroup cues strengthen support for violence and anti-Muslim policies.* University of Michigan, Department of Statistics.
Jiang, R., Chiappa, S., Lattimore, T., György, A., & Kohli, P. (2019). Degenerate feedback loops in recommender systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 383–390).
Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 133–142).
Johnson, N. F., Velásquez, N., Restrepo, N. J., Leahy, R., Gabriel, N., El Oud, S., … Lupu, Y. (2020). The online competition between pro-and anti-vaccination views. *Nature*, *582*(7811), 230–233.
Kilgo, D. K., Harlow, S., García-Perdomo, V., & Salaverría, R. (2018). A new sensation? An international exploration of sensationalism and social media recommendations in online news publications. *Journalism*, *19*(11), 1497–1516.
Kilgo, D. K., & Sinta, V. (2016). Six things you didn't know about headline writing: Sensationalistic form in viral news content from traditional and digitally native news organizations. *Quieting the Commenters: The Spiral of Silence's Persistent Effect*, *111*.
Kiritchenko, S., & Nejadgholi, I. (2020). Towards ethics by design in online abusive content detection. *arXiv preprint arXiv:2010.14952*.
Kiritchenko, S., Nejadgholi, I., & Fraser, K. C. (2021). Confronting abusive language online: A survey from the ethical and human rights perspective. *Journal of Artificial Intelligence Research*, *71*, 431–478.
Klandermans, B., & De Weerd, M. (2000). Group identification and political protest. *Self, Identity, and Social Movements*, *13*, 68–90.
Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, *42*(8), 30–37.
Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788–8790.
Kteily, N., Bruneau, E., Waytz, A., & Cotterill, S. (2015). The ascent of man: Theoretical and empirical

evidence for blatant dehumanization. *Journal of Personality and Social Psychology*, *109*(5), 901.
Lavis, A., & Winter, R. (2020). # Online harms or benefits? An ethnographic analysis of the positives and negatives of peer-support around self-harm on social media. *Journal of Child Psychology and Psychiatry*, *61*(8), 842–854.
Ledwich, M., & Zaitsev, A. (2019). Algorithmic extremism: Examining YouTube's rabbit hole of radicalization. *arXiv preprint arXiv:1912.11211*.
Levy, R. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review*, *111*(3), 831–70.
Lingiardi, V., Carone, N., Semeraro, G., Musto, C., D'Amico, M., & Brena, S. (2020). Mapping Twitter hate speech towards social and sexual minorities: a lexicon-based approach to semantic content analysis. *Behaviour & Information Technology*, *39*(7), 711–721.
Loewenstein, J. (2019). Surprise, recipes for surprise, and social influence. *Topics in Cognitive Science*, *11*(1), 178–193.
Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K., & Larson, H. J. (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour*, *5*(3), 337–348.
Martin, A. M. (2008). Hope and exploitation. *Hastings Center Report*, *38*(5), 49–55.
McCauley, C., & Moskalenko, S. (2008). Mechanisms of political radicalization: Pathways toward terrorism. *Terrorism and Political Violence*, *20*(3), 415–433.
McKay, S., & Tenove, C. (2021). Disinformation as a threat to deliberative democracy. *Political Research Quarterly*, *74*(3:703-717).
McNiel, D. E., & Binder, R. L. (2007). Effectiveness of a mental health court in reducing criminal recidivism and violence. *American Journal of Psychiatry*, *164*(9), 1395–1403.
McWhirter, R. E., Critchley, C. R., Nicol, D., Chalmers, D., Whitton, T., Otlowski, M., & ...Dickinson, J. L. (2014). Community engagement for big epidemiology: deliberative democracy as a tool. *Journal of Personalized Medicine*, *4*(4:459-474).
Meadows, D. (2008). Thinking in systems: a primer. *White River Junction: Chelsea Green Publishing*.
Miller, T. (2019). Explanation in Artificial Intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*(1-38).
Moghaddam, F. M. (2005). The staircase to terrorism: A psychological exploration. *American psychologist*, *60*(2), 161.
Mondal, M., Silva, L. A., & Benevenuto, F. (2017). A measurement study of hate speech in social media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media* (pp. 85–94).
Monsees, L. (2021). Information disorder, fake news and the future of democracy. *Globalizations*(1-16).
Moretti, F. (2013). Distant reading. *Verso*.
Mosseri, A. (2018). *Helping ensure news on Facebook is from trusted sources.* Facebook Newsroom.
Mourão, R. R., & Robertson, C. T. (2019). Fake news as discursive integration: An analysis of sites that publish false, misleading, hyperpartisan and sensational information. *Journalism Studies*, *20*(14), 2077–2095.
MRX, D. (2018). Conspiracy theories and philosophy: bringing the epistemology of a freighted term into the social sciences. *Conspiracy Theories and the People Who Believe Them, Oxford University Press*.
Mueller, M. (2012). Scalable reading. *Northwestern University*.
Müller, K., & Schwarz, C. (2021). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, *19*(4), 2131–2167.
Munger, K., & Phillips, J. (2019). A supply and demand framework for YouTube politics. *Penn State, University Park*.
Munn, L. (2021). *Alt-Right Pipeline.* Manuscript, Western Sydney University.
Nadar, S. (2014). Stories are data with soul. *Agenda*, *28*(1).
Naumov, M., Mudigere, D., Shi, H.-J. M., Huang, J., Sundaraman, N., Park, J., ... others (2019). Deep learning recommendation model for personalization and recommendation systems. *arXiv preprint arXiv:1906.00091*.
Neumann, P. R. (2003). The trouble with radicalization. *International affairs*, *89*(4), 873–893.
Newton, K. (2018). White noise: Some Aucklanders have more say in their city's future than others. Retrieved from https://shorthand.radionz.co.nz/white-noise/.
Ngata, T. (2020). The rise of Māori MAGA. *E-Tangata*.
OECD. (2020). *Current approaches to terrorist and violent extremist content among the global top 50 online*

content-sharing services. OECD Digital Economy Papers 296.
Oldfield, L. D. (2016). Vested interest and public action in a climate of participatory democracy: Water fluoridation decisions in Australasian local government. *Masters thesis, University of Waikato*.
Oshikawa, R., Qian, J., & Wang, W. Y. (2020). A survey on natural language processing for fake news detection. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)* (pp. 6086–6093).
Paasch-Colberg, S., Strippel, C., Trebbe, J., & Emmer, M. (2021). From insult to hate speech: Mapping offensive language in German user comments on immigration. *Media and Communication*, *9*(1), 171–180.
Padín, P. F., González-Rodríguez, R., Verde-Diego, C., & Vázquez-Pérez, R. (2021). Social media and eating disorder psychopathology: A systematic review. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *15*(3).
Papadamou, K., Zannettou, S., Blackburn, J., De Cristofaro, E., Stringhini, G., & Sirivianos, M. (2021). "How over is it?" Understanding the Incel Community on YouTube. *arXiv preprint arXiv:2001.08293*.
Pariser, E. (2011). *The filter bubble: What the internet is hiding from you*. Penguin UK.
Pariser, E. (2015). *A new study from facebook reveals just how much it filters what you see.* Gizmodo.
Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, *3*, 96–146.
Peralta, A. F., Neri, M., Kertész, J., & Iñiguez, G. (2021). *The effect of algorithmic bias and network structure on coexistence, consensus, and polarization of opinions.*
Perliger, A., Koehler-Derrick, G., & Pedahzur, A. (2016). The gap between participation and violence: Why we need to disaggregate terrorist 'profiles'. *International Studies Quarterly*, *60*(2), 220–229.
Perra, N., & Rocha, L. (2019). Modelling opinion dynamics in the age of algorithmic personalisation. *Scientific Reports*, *9*.
Persily, N., & Tucker, J. A. (2020). *Social media and democracy: The state of the field, prospects for reform*. Cambridge University Press.
Post, J. M. (2007). *The mind of the terrorist: The psychology of terrorism from the IRA to al-Qaeda*. St. Martin's Press.
Prior, M. (2007). *Post-broadcast democracy: How media choice increases inequality in political involvement and polarizes elections*. Cambridge University Press.
Rae, J. A. (2012). Will it ever be possible to profile the terrorist? *Journal of Terrorism Research*, *3*(2), 64–74.
Rathje, S., Van Bavel, J. J., & van der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, *118*(26).
Ressa, M. (2016). *Propaganda War: Weaponizing the Internet.* `rappler.com`.
Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A., & Meira Jr, W. (2020). Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 131–141).
Ricci, F., Rokach, L., & Shapira, B. (2015). Recommender systems: introduction and challenges. In *Recommender Systems Handbook* (pp. 1–34). Springer.
Robertson, A. (2020). *Facebook will add anti-misinformation posts to your News Feed if you liked fake coronavirus news.* The Verge.
Romano, A. (2020). *New Yahoo News/YouGov poll shows coronavirus conspiracy theories spreading on the right may hamper vaccine efforts.* Yahoo News.
Rony, M. M. U., Hassan, N., & Yousuf, M. (2017). Diving deep into clickbaits: Who use them to what extents in which topics with what effects? In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017* (pp. 232–239).
Rossi, W. S., Polderman, J., & Frasca, P. (2018). The closed loop between opinion formation and personalised recommendations. *ArXiv*, *abs/1809.04644*.
Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. (2020). Hatecheck: Functional tests for hate speech detection models. *arXiv preprint arXiv:2012.15606*.
Rychwalska, A., & Roszczyńska-Kurasińska, M. (2018). Polarization on social media: when group dynamics leads to societal divides. In *Proceedings of the 51st Hawaii International Conference on System Sciences.*
Salkind, N. (2010). The encyclopedia of research design. *Sage*.
Salminen, J., Almerekhi, H., Milenković, M., Jung, S.-g., An, J., Kwak, H., & Jansen, B. J. (2018). Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *Twelfth International AAAI Conference on Web and Social Media.*

Sandover, R., Moseley, A., & Devine-Wright, P. (2021). Contrasting views of citizens' assemblies: Stakeholder perceptions of public deliberation on climate change. *Politics and Governance*, *9*(2:76-86).
Schwartz, S. J., Dunkel, C. S., & Waterman, A. S. (2009). Terrorism: An identity theory perspective. *Studies in Conflict & Terrorism*, *32*(6), 537–559.
Sellars, A. (2016). *Defining hate speech.* Berkman Klein Center Research.
Shani, G., & Gunawardana, A. (2011). Evaluating recommendation systems. In *Recommender Systems Handbook* (pp. 257–297). Springer.
Shearer, E., & Mitchell, A. (2021). *News use across social media platforms in 2020.* Pew Research Center.
Shermer, M. (2011). *The believing brain: From ghosts and gods to politics and conspiracies—how we construct beliefs and reinforce them as truths*. Macmillan.
Silverman, C. (2016). *This analysis shows how viral fake election news stories outperformed real news on Facebook.* Buzzfeed, https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook.
Simandan, D. (2020). Being surprised and surprising ourselves: A geography of personal and social change. *Progress in Human Geography*, *44*(1), 99–118.
Sinpeng, A., Gueorguiev, D., & Arugay, A. A. (2020). Strong fans, weak campaigns: Social media and Duterte in the 2016 Philippine election. *Journal of East Asian Studies*, *20*(3), 353–374.
Sîrbu, A., Pedreschi, D., Giannotti, F., & Kertész, J. (2019). Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model. *PloS one*, *14*(3), e0213246.
Siu, A. (2017). Deliberation & the challenge of inequality. *Daedalus*, *146*(3:119-128).
Sivaneswaran, S., Chong, G. T., & Blinkhorn, A. S. (2010). Successful fluoride plebiscite in the township of Deniliquin, New South Wales, Australia. *Journal of Public Health Dentistry*, *70*(2: 163-166).
Smith, A. (2009). *Radicalization—a guide for the perplexed.* National Security Criminal Investigations, Trans., Royal Canadian Mounted Police.
Smith, L. T. (2012). Decolonizing methodologies: research and indigenous peoples. *Zed Books, London*.
Smith, N., & Graham, T. (2019). Mapping the anti-vaccination movement on Facebook. *Information, Communication & Society*, *22*(9), 1310–1327.
Soar, M., Louise-smith, V., Dentith, M., Barnett, D., Hannah, K., Valentino Dalla Riva, G., & Sprole, A. (2020). Evaluating the infodemic: assessing the prevalence and nature of COVID-19 unreliable and untrustworthy information in Aotearoa New Zealand's social media. *Working paper. https://www.tepunahamatatini.ac.nz/covid-19/*.
Steinhardt, J. (2021). *How much do recommender systems drive polarization?* Blog post, https://jsteinhardt.stat.berkeley.edu/blog/recsys-deepdive.
Stewart, A. J., McCarty, N., & Bryson, J. J. (2020). Polarization under rising inequality and economic decline. *Science advances*, *6*(50), eabd4201.
Stray, J. (2020). Aligning AI optimization to community well-being. *International Journal of Community Well-Being*, *3*(4), 443–463.
Suiter, J. (2016). Post-truth politics. *Political Insight*, *7*(3:25-27).
Summers, N. (2020). *Facebook rolls back News Feed change that prioritized mainstream media.* Engadget post.
Sunstein, C. (2001). *Republic.com*. Princeton University Press.
Sunstein, C. R. (2000). Deliberative trouble? Why groups go to extremes. *The Yale Law Journal*, *110*(1), 71–119.
Sutton, R. M., & Douglas, K. M. (2020). Conspiracy theories and the conspiracy mindset: Implications for political ideology. *Current Opinion in Behavioral Sciences*, *34*, 118–122.
Syed-Abdul, S., Fernandez-Luque, L., Jian, W.-S., Li, Y.-C., Crain, S., Hsu, M.-H., ... others (2013). Misleading health-related information promoted through video-based social media: anorexia on youtube. *Journal of medical Internet research*, *15*(2), e30.
Tait, A. (2017). *Spitting out the red pill: Former misogynists reveal how they were radicalised online.* New Statesman.
Tajfel, H., Turner, J. C., Austin, W. G., & Worchel, S. (1979). An integrative theory of intergroup conflict. *Organizational identity: A reader*, *56*(65), 9780203505984–16.
Talbot, H., & Alaili, N. (2021). The Edge of the Infodemic: Challenging Misinformation in Aotearoa. *Te Mana Whakaatu– The Classification Office, Wellington*.
Tenenboim, O., & Cohen, A. A. (2015). What prompts users to click and comment: A longitudinal study of

online news. *Journalism*, *16*(2), 198–217.
Tenove, C. (2020). Protecting democracy from disinformation: Normative threats and policy responses. *The International Journal of Press/Politics*, *25*(3:517-537).
Tierney, S. (2008). Creating communities in cyberspace: pro-anorexia web sites and social capital. *Journal of Psychiatric and Mental Health Nursing*, *15*(4), 340–343.
Timmins, B. (2021). *Twitter works with news sites to tackle disinformation.* BBC News.
Tromble, R. (2021). Where have all the data gone? A critical reflection on academic digital research in the post-API age. *Social Media+ Society*, *7*(1), 2056305121988929.
Twitter. (2019). *Q3 2019 Letter to Shareholders.* Twitter Inc.
Van Bavel, J. J., Harris, E. A., Pärnamets, P., Rathje, S., Doell, K., & Tucker, J. A. (2020). Political psychology in the digital (mis) information age: A model of news belief and sharing. *Social Issues and Policy Review*, *15*(1), 84–113.
Van Stekelenburg, J. (2014). Going all the way: Politicizing, polarizing, and radicalizing identity offline and online. *Sociology Compass*, *8*(5), 540–555.
Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., & Margetts, H. (2019). Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online* (pp. 80–93).
Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.
Walters, L. (2021). Kiwis more vulnerable to online extremism in lockdown. *Stuff.co.nz*.
Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe*, *27*.
Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*.
Watson, A., & Mace, L. (2014). Drinking water fluoridation in New Zealand. Retrieved from https://www.waternz.org.nz/.
Whitehead, A. L., & Perry, S. L. (2020). How culture wars delay herd immunity: Christian nationalism and anti-vaccine attitudes. *Socius*, *6*, 2378023120977727.
Whittaker, J., Looney, S., Reed, A., & Votta, F. (2021). Recommender systems and the amplification of extremist content. *Internet Policy Review*, *10*(2), 1–29.
Wiegand, M., Ruppenhofer, J., Schmidt, A., & Greenberg, C. (2018). Inducing a lexicon of abusive words—a feature-based approach. In *Proceedings of the Association for Computational Linguistics.*
Wikforss, C. (2020). Deliberative democracy could be used to combat fake news–but only if it operates offline. *Democratic Audit Blog*.
Williams, M. L., Burnap, P., Javed, A., Liu, H., & Ozalp, S. (2020). Hate in the machine: Anti-Black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, *60*(1), 93–117.
Winstanley, A. (2005). The not-so-hidden politics of fluoridation. *Policy & politics*, *33*(3:367-385).
Wyman, R. A., Mahoney, E. K., & Børsting, T. (2015). Community water fluoridation: attitudes and opinions from the New Zealand Oral Health Survey. *Australian and New Zealand Journal of Public Health*.
Yoshida, M., Sakaki, T., Kobayashi, T., & Toriumi, F. (2021). Japanese conservative messages propagate to moderate users better than their liberal counterparts on Twitter. *Nature Scientific Reports*, *11*, 19224.
YouTube. (2019). *The Four Rs of Responsibility, Part 2: Raising authoritative content and reducing borderline content and harmful misinformation.* https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce/.
Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). SemEval-2019 Task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 75–86).
Zannettou, S., Finkelstein, J., Bradlyn, B., & Blackburn, J. (2020). A quantitative approach to understanding online antisemitism. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 14, pp. 786–797).
Zhang, S., Chen, H., Ming, X., Cui, L., Yin, H., & Xu, G. (2021). Where are we in embedding spaces? In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (pp. 2223–2231).
Zhao, Z., Hong, L., Wei, L., Chen, J., Nath, A., Andrews, S., ... Chi, E. (2019). Recommending what video to

watch next: a multitask ranking system. In *Proceedings of the 13th ACM Conference on Recommender Systems* (pp. 43–51).

Zhou, L. (2015). A survey on contextual multi-armed bandits. *arXiv preprint arXiv:1508.03326*.

Zickmund, S. (1997). Approaching the radical other: The discursive culture of cyberhate. *Virtual culture: Identity and communication in cybersociety*, 185–205.