Transparency Mechanisms for Social Media Recommender Algorithms: From Proposals to Action

Tracking GPAI's Proposed Fact Finding Study in This Year's Regulatory Discussions

November 2022

GPAI | THE GLOBAL PARTNERSHIP ON ARTIFICIAL INTELLIGENCE

This report was developed by Experts and Specialists involved in the Global Partnership on Artificial Intelligence's project on Responsible AI for Social Media Governance. The report reflects the personal opinions of the GPAI Experts and Specialists involved and does not necessarily reflect the views of the Experts' organisations, GPAI, or GPAI Members. GPAI is a separate entity from the OECD and accordingly, the opinions expressed and arguments employed therein do not reflect the views of the OECD or its Members.

Acknowledgements

This report was developed in the context of the Responsible AI for Social Media Governance project, with the steering of the project Co-Leads, supported by the GPAI Responsible AI Working Group. The GPAI Responsible AI Working Group agreed to declassify this report and make it publicly available.

Co-leads:
Alistair Knott, School of Engineering and Computer Science, Victoria University of Wellington
Dino Pedreschi, Department of Computer Science, University of Pisa

The report was written by: **Alistair Knott**[*], School of Engineering and Computer Science, Victoria University of Wellington; **Dino Pedreschi**[*], Department of Computer Science, University of Pisa; **Tapabrata Chakraborti**[†], University of Oxford and the Alan Turing Institute; **David Eyers**[†], Department of Computer Science, University of Otago; **Raja Chatila**[*], Sorbonne University; **Andrew Trotman**[†], Department of Computer Science, University of Otago; **Ricardo Baeza-Yates**[**], Institute for Experiential AI, Northeastern University; **Lama Saouma**[†], GPAI's Montreal Center of Expertise - CEIMIA; **Virginia Morini**[†], Istituto di Scienza e Tecnologie dell'Informazione, NIRC; **Valentina Pansanella**[†], Scuola Normale Superiore, University of Pisa.

GPAI would like to thank the following people who gave valuable feedback as the project progressed, and commented on drafts of this report: Toshiya Jitsuzumi[*], from Chuo University's Faculty of Policy Studies, Hector Selby[†] and Gagandeep Bhandal[†], from the UK Home Office's Online Policy Unit, David Reid[†] and Paul Ash[†], from New Zealand's Department of the Prime Minister and Cabinet, members of the Recommender Systems project coordinated by GIFCT's Technical Approaches working group led by Tom Thorley[†], Jonathan Stray[†] from Berkeley's Center for Human-Compatible AI, Gillian Hadfield[†] from Toronto's Schwartz Reisman Institute, and Chris Meserole[†], from the Brookings Institute. Also, thanks to the International Centre of Expertise in Montréal for Artificial Intelligence (CEIMIA) for their support.

[*] Expert of GPAI's Responsible AI Working Group
[**] Observer at GPAI's Responsible AI Working Group
[†] Invited specialist

Citation

GPAI 2022. *Transparency Mechanisms for Social Media Recommender Algorithms: From Proposals to Action. Tracking GPAI's Proposed Fact Finding Study in This Year's Regulatory Discussions.* Report, November 2022, Global Partnership on AI.

Contents

1

Executive summary 1
1 Introduction 3
1.1 Recommender systems: an important topic for GPAI attention 3
1.2 A prima facie cause for concern with recommender systems 4
1.3 A focus on Terrorist and Violent Extremist Content 4
1.4 Structure of this report 5
1.5 A note about the target readership for the report 6
2 Do recommender systems have effects on users' attitudes towards TVEC? GPAI's proposed fact-finding study 7
2.1 A survey of 'external' methods for studying recommender system effects . . . 7
2.2 A survey of 'internal' methods for studying recommender system effects . . . 10
2.3 GPAI's proposed fact-finding studies 15
2.4 Summary 18
3 Discussions about GPAI's fact-finding study, and related transparency initiatives 20
3.1 Discussions between GPAI and Twitter 20
3.2 Discussions at GIFCT 21
3.3 Discussions with government groups 28
3.4 Discussions in the Christchurch Call's Algorithms workstream 28
3.5 A survey of other work on recommender system transparency and functionality 32
3.6 A survey of regulatory initiatives involving recommender systems 36
4 Thoughts on how to streamline discussions about recommender system transparency 41
4.1 Create high-level support for transparency initiatives 41
4.2 Improve awareness about the processes of government amongst GPAI experts 42
4.3 Involve company engineers in transparency discussions 42
4.4 Focus discussions on concrete pilot studies 43
4.5 Better interactions between cooperative and regulatory discussions 43
4.6 Create a public science around recommender system effects 44
Bibliography 45

Executive summary

Social media platforms rely on several kinds of AI technology for their operation. Much of the appeal of social media platforms comes from their ability to deliver content that is *tailored* to individual users. This ability is provided in large part by AI systems called **recommender systems**: these systems are the focus of our project.

Recommender systems curate the 'content feeds' of platform users, using machine learning techniques to tailor each user's feed to the kinds of item they have engaged with in the past. They essentially function as a personalised newspaper editor for each user, choosing which items to present, and which to withhold. They rank amongst the most pervasive and influential AI systems in the world today.

The starting point for our project is a concern that recommender systems may lead users in the direction of harmful content of various kinds. This concern is at origin a technical one, relating to the AI methods through which recommender systems learn. But it is also a social and political one, because the effects of recommender systems on platform users could potentially have a significant influence on currents of political opinion.

At present, there is very little public information about the effects of recommender systems on platform users: we know very little about how information is disseminated to users on social media platforms. It is vital that governments, and the public, have more information about how recommender systems steer content to platform users, particularly in domains of harmful content.

In the first phase of our project, we reviewed possible methods for studying the effects of recommender systems on user platform behaviour. We concluded the best methods available for studying these effects are the methods that companies use themselves. These are methods that are only available internally to companies. We proposed *transparency mechanisms*, in which these company-internal methods are used to address questions in the public interest, about possible harmful effects of recommender systems.

We focussed on the domain of **Terrorist and Violent Extremist Content** (**TVEC**), because this type of content is already the focus of discussion in several ongoing initiatives involving companies, including the Global International Forum to Counter Terrorism (GIFCT) and the Christchurch Call to Eliminate TVEC Online. Our proposal was for a form of **fact-finding study**, that we argued would surface relevant information about recommender system effects in this area, without compromising the rights of platform users, or the intellectual property of companies. We presented and argued for this proposed fact-finding study at last year's GPAI Summit.

Over the past year, our project has pursued the practical goal of piloting our proposed fact-finding study in one or more social media companies. This has involved discussions with several companies, often mediated by governments; and participation in several international initiatives relating to TVEC, in particular the Christchurch Call and GIFCT. At the recent Christchurch Call Summit, a scheme for running a pilot project of the kind we advocate was announced: the initiative involves two governments (the US and New Zealand) and two tech companies (Twitter and Microsoft), and centres on the trialling of 'privacy-enhancing technologies' developed by a third organisation, OpenMined. In this report, we will summarise the discussions that led to this initiative, in the context of other ongoing discussions around transparency mechanisms for recommender systems.

We are very much looking forward to participating in the scheme initiated by the US and New Zealand, and engaging with Twitter, Microsoft and OpenMined. But discussions about recommender system transparency will certainly continue beyond this initiative as well.[1] Our report also offers some recommendations about how these ongoing discussions can be made more efficient.

- First, we suggest that discussions should have more involvement from *company engineers*. At present, companies are represented primarily through legal and policy teams. But the questions under discussion concern technical mechanisms operating within companies: the engineers who design and use these mechanisms could make valuable contributions, under suitable non-disclosure arrangements.
- Second, while privacy-enhancing technologies are potentially very valuable, we foresee the need for an *ongoing discussion* about these technologies between companies and external stakeholders. As companies' technologies develop, and new questions arise, the functionality of these technologies may have to be expanded, or adjusted.
- Third—and on a related note—we don't see privacy-enhancing technologies as a substitute for discussions with company engineers. We think discussions with engineers should begin immediately, while privacy-enhancing technologies are being developed; and we foresee a role for discussions with engineers even after these technologies are first put in place.
- Fourth, we suggest transparency discussions about recommender systems should focus on the specification and implementation of *pilot projects* that can be trialled in particular companies. Discussions with company engineers can define a set of possible pilot projects. The proposed pilots can then provide a concrete focus for discussions with company lawyers, to ensure protection of user rights and company intellectual property. These discussions with individual companies can inform broader discussions about general transparency processes that apply across companies.
- Fifth, we foresee an ongoing role for the piloting of recommender system transparency mechanisms, as social media platforms continue to change and develop. We suggest that discussions about pilots could be coordinated by an independent regulatory body, sitting between companies and governments, and informed by a new public science of recommender system effects.

[1] Elon Musk's acquisition of Twitter was announced just as our report was going to press. But as noted here, our project involves a broad range of discussions, that go well beyond the Twitter initiative.

1 Introduction

1.1 Recommender systems: an important topic for GPAI attention

The distinctive appeal of social media systems stems in large part from their ability to provide content that is *tailored* to individual users. Much of this ability is implemented by AI systems called **recommender systems**, that deliver personalised content feeds to each user.[1] Given that more than half the world's population are social media users, and that Internet users spend over two and a half hours a day on average on social media sites (see Kemp, 2022 for indicative data on both points), recommender systems must be counted among the most pervasive and influential AI systems in today's world: they are therefore a very natural focus of attention for GPAI.

Of particular interest for our working group, on Responsible Use of AI, is that fact that Internet users are increasingly consuming their *political news and current affairs* from social media sites (see Watson, 2022 for indicative data). Social media recommender systems essentially function as personalised newspaper editors for billions of social media users around the world: in this role, they have a huge influence on the dynamics of public and political opinion. It is of great importance to ensure that social media companies are acting *responsibly* in exercising this influence.

Responsible journalism is a fairly well defined concept as it arises in discussions of conventional media. The editors of news content for newspapers or TV/radio programmes have some clear legal responsibilities (for instance, to avoid libel, graphic content), and some social responsibilities that go beyond these (for instance, to provide 'balanced coverage', to avoid 'misinformation') that are less widely enforced, but still relatively well understood. What it means for a social media recommender system to be 'responsible' in its selection of news for users is much less well understood, and is an important new topic for research and discussion.

To be clear, our GPAI project won't be suggesting that companies should have *legal responsibility* for the content disseminated by recommender systems. There are well known difficulties with that proposal, which we will briefly touch on. Our project aims to contribute to a broader discussion of responsibility in two basic ways, that both draw on our expertise as AI practitioners. Firstly, we want to identify the best way to *study* and *measure* the effects of recommender systems on the experience of platform users. These effects are often discussed in the media, but in fact there is very little data about them in the public domain. We believe the first step in a discussion about recommender algorithms is to surface *better information* about what effects they have (or do not have) on users, for policymakers, citizens' groups, and other stakeholders. We have studied the different methods that are available for providing this information, and we have made a proposal about the method which is best.

Secondly, we argue there is a specific concern about the effects of recommender systems on platform users that deserves particular attention. This concern arises from the way recommender systems *learn*, which again as AI experts we are well placed to discuss. We will articulate the concern in the next section.

[1] Recommender systems are also known as **recommender algorithms** and **content-sharing algorithms**; we'll use these terms interchangeably.

1.2 A prima facie cause for concern with recommender systems

A recommender system learns about each individual platform user by observing their *behaviour* on the platform: that is, how they respond to the content items they encounter (see e.g. Ricci *et al.*, 2015; Lada *et al.*, 2021). The user might click on an item, or 'like' it, or comment on it, or forward it, or ignore it. Through a variety of AI methods, the recommender system learns about the *kinds* of item the user engages with: it then prioritises items of these same kinds in the user's subsequent content feed. That's a simple statement of what recommender systems do. But it's enough to express the concern we focus on in our project.

The basic concern is that because a recommender system *chooses* the order of items in each user's feed, it also *influences its own subsequent learning*. A recommender system learns about its users on an *ongoing basis*, regularly making new observations about their behaviour, and updating its model of what each user engages with. But what the system observes at any given point are the user's responses (clicks, likes etc.) *to the items the system recommended for them*, based on its earlier model of that user.[2]

The fact that recommender systems simultaneously *learn from*, and *influence* users' behaviour means they may learn to *reinforce their existing user models*, giving each user more content of the kind they already recommended. The concern, stated more specifically, is that recommender systems may lead users towards progressively narrower domains of content, in directions they might not otherwise have travelled. This concern can be readily shown in theoretical studies of recommender systems: an analysis by Google DeepMind (Jiang *et al.*, 2019) shows the effect clearly, and our previous report (GPAI, 2021) explains the effect in detail.

1.3 A focus on Terrorist and Violent Extremist Content

In our project, we specifically aim to study whether recommender systems have any tendency to move users towards engaging with *extremist content*. The journey towards extremism has been studied in many ways; again, we review this body of work in our previous report. The key issue for recommender systems, which exacerbates the concern mentioned above, is that social media users are known to show small *biases* towards extreme content of various kinds, that act as another influence on the content items they engage with. For instance, they have a tendency to share political messages containing 'moral emotional expressions' (Brady *et al.*, 2017; Brady and van Bavel, 2021), particularly negative ones (Crockett, 2017; Brady and van Bavel, 2021; 2021), messages that refer to a political 'out-group' (Rajthe *et al.*, 2021), and messages that contain falsehoods (Vosoughi *et al.*, 2018). If these biases *persist* while a user interacts with a recommender system, the system's repeated updates of its user model may lead the user towards messages containing increasing levels of negative political emotions, an increasing focus on political out-groups, and increasing amounts of misinformation—and potentially towards domains of violent extremism. Again, our earlier report (GPAI, 2021) presents these concerns and the studies that support them in detail.

At this point, our *technical* concerns about learning in recommender systems connect with very active *social* and *political* concerns about content on social media. There is widespread concern about the proliferation of harmful content on social media platforms. This content can be of many kinds—but much recent discussion centres around 'Terrorist and Violent Extremist Content' ('TVEC'). In the Global Internet Forum to Counter Terrorism (GIFCT), tech companies are collaborating to share information about content of this kind, and to

[2] It is well known that users preferentially engage with items that are presented early in a list of candidates; see e.g. Joachims *et al.*, 2005.

develop protocols for identifying it and removing it, with support from governments and citizens' groups. These efforts are extended by work coordinated by the Christchurch Call to Eliminate TVEC Online, which also provides a forum for collaborations between companies, governments and other stakeholders. We chose to focus our project on TVEC so as to align productively with these existing groupings.

Recall that the broad objective of our study is to provide the public with good information about the effects of social media recommender systems on users. Our main concern is with *methods*: we want to identify the best method for studying effects on users. The specific question we want to explore is *whether recommender systems have any influence on users' attitudes towards TVEC*. A focus on TVEC is of interest to us technically, as AI theorists, because there are technical concerns recommender systems may lead users in the direction of 'extreme' content of various kinds. But it also connects to pressing ongoing discussions about harmful online content. The topic of TVEC thus recommends itself particularly for attention by GPAI.

1.4 Structure of this report

In the first phase of our project, culminating in our 2021 report (GPAI, 2021), we reviewed possible methods for studying the effects of social media recommender systems on users. This was largely a question of scientific methodologies: how can we best get the information that stakeholders need? We concluded that the best way to get this information was to use methods that are only available *within social media companies*. Companies have sophisticated ways at their disposal for studying the effects of recommender systems on users. In fact, all companies that we surveyed are *already* studying the effects of their recommender systems on their users, for a variety of purposes, and using a variety of methods. In our initial report, our central proposal was that these company-internal methods could be *co-opted* to address our key research question, about the effects of recommender systems on user attitudes towards TVEC. We proposed a family of studies—that we called 'fact-finding studies'—that could be safely conducted within companies to surface information about these effects, in ways that (we argued) would not compromise the privacy of platform users, or the intellectual property of companies. Effectively, we proposed mechanisms that would provide *transparency* about the effects of recommender systems on users. Our proposal thus connects with ongoing discussions about transparency processes for tech companies, which are intimately linked with discussions about harmful online content. We will briefly summarise our proposal, and the arguments that motivated it, in Chapter 2.

Crucially, the particular method we advocate for studying recommender system effects can only be implemented *within social media companies*, using processes that are only available to them. This means our fact-finding study necessarily requires *engagement with companies*. In the past year, we have participated in many discussions, with social media companies and with governments and citizens' groups, about how to manage this engagement. In Chapter 3 we will summarise these discussions, and situate them in the broader context of discussions that are currently under way, in company initiatives, academic projects and legislatory processes. Our aim here is partly to describe what we have done this year—and in particular, to describe the processes that led up to the US-New Zealand initiative involving Twitter, Microsoft and OpenMined. But we also aim to present a broader picture of the complex structures that are emerging internationally, to support ongoing discussions in the area of recommender system transparency. We hope this picture will be helpful for those involved in organising and participating in these discussions. We think it also offers some suggestions about how these discussions could be improved—and indeed, for changes that could be made within GPAI's own processes. We conclude in Chapter 4 by outlining these suggestions.

1.5 A note about the target readership for the report

Our first report (GPAI, 2021) assumed a technical audience at certain points. The current report does not assume any specialist knowledge of AI, though we do give references for those who would like further technical details: we introduce the relevant technical ideas very informally. The target audience for the current report are the people who participate in discussions about tech transparency, including representatives from companies, governments, and civil society groups, as well as technical specialists in AI and Computer Science. Our focus is squarely on how to move from technical proposals about recommender system transparency to practical actions.

2 Do recommender systems have effects on users' attitudes towards TVEC? GPAI's proposed fact-finding study

In the first phase of our project, we surveyed many different methods for studying the effects of recommender systems on users. We divided these into 'external methods' that could be used outside social media companies, drawing on publicly available information, or data gathered from user experiments, and 'internal methods', that are only available within social media companies. We'll briefly review external methods in Section 2.1 and internal methods in Section 2.2.

We identified many shortcomings with external methods, and found that internal methods deployed within companies provide far better information about recommender system effects on users. Moreover, the one study using internal methods that has been published so far (Huszár *et al.*, 2022) provides clear evidence that recommender systems strongly influence users' consumption of political content, as we'll discuss in Section 2.2.2. Accordingly, we proposed that *companies should collaborate with external stakeholders*, to design and conduct studies that provide the information about recommender system effects that is currently lacking in public and policy discussions.

Our proposal centred on studies investigating whether recommender systems have effects on users' attitudes towards TVEC—a particularly pressing question, for the reasons discussed in Section 1.3. We called these collaborative company-internal studies **fact-finding studies**. We proposed a family of methods that could be used in a fact-finding study, and we made some suggestions about how companies and external stakeholders could come together to design such a study. We will summarise these proposals in Section 2.3.

2.1 A survey of 'external' methods for studying recommender system effects

Academic researchers have found a variety of ways to study the effects of recommender systems 'from the outside', without special access to company processes and data. We will review the main methodologies that have been used, illustrating each with a study that examines effects relevant to TVEC. A longer review with more examples can be found in Knott *et al.* (2021) Ch3; see also Whittaker (2022).

Population studies use data about large groups of people, juxtaposing data about TVEC-related attitudes with data about social media use. For instance, Boxell *et al.* (2017) surveyed political polarisation and social media use in the US from 1996 to 2016. They find that polarisation has increased the most amongst those aged over 65—but also that this age group engaged less with social media than other age groups over the same time period. If social media use is a primary cause of polarisation, they argue we would expect less polarisation in those over 65; this evidence suggests that social media use is not a primary cause of polarisation.

Population studies provide useful information—in particular, about differences in effects of social media over place, time and demographic groups. But they also have methodological

shortcomings. In particular, there are often *confounding variables* that could explain differences over groups but did not feature in the analysis. In relation to Boxell *et al.*'s study, for instance, it is also relevant that older people are more trusting of material they encounter online than younger people (see e.g. Brashier and Schacter, 2020). It is hard to account for all possible confounding variables in a population study.

Browser logging studies observe the behaviour of volunteers on social media sites over a period of time. These studies can focus in more detail on recommender system effects, because they can observe how users interact with the content that recommender systems offer them. For instance, Flaxman *et al.* (2016) examined the web-browsing behaviour of US-based users who had volunteered to share their web browsing history for research purposes. Flaxman *et al.* found that nearly all the news seen by these users came from visits to their favourite news outlet; much less came from following links offered by social media recommender systems. 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 provides some evidence that social media contributes to political polarisation. A recent browser-logging study of YouTube by Brown *et al.* (2022) found a similar effect: YouTube's recommender algorithm channeled US users into 'mild echo chambers' separating liberals from conservatives. Other browser logging studies have found the opposite effect, however. For instance, Scharkow *et al.* (2020) found that users in Germany who browsed social media sites were more likely to have a 'larger and more varied' news diet than users who did not.

Browser logging studies surface a rich picture of how social media sites are used, at different times and places. Of course, we expect differences in logs sampled from different populations. But browser logging studies may also suffer from *systematic* problems with sampling volunteers. Perhaps users who are more susceptible to radicalisation are also less likely to share their browsing data with experimental researchers. Note also that consumption of news on social media sites has grown since Flaxman *et al.*'s study; in 2020, more than half of US adults said they get news from social media 'often' or 'sometimes' (Shearer and Mitchell, 2021). And even if users only sometimes use recommender algorithms to access current affairs content, they may still be subject to the cumulative effects discussed in Sections 1.2 and 1.3.

Public API studies make use of data surfaced publicly by social media companies about activity on their platform. For instance, Ledwich and Zaitsev (2019) examine data about recommendations made available by YouTube's API, that allows a researcher to query how often a user watching a given video A is recommended another video B. Ledwich and Zaitsev examine videos from 800 YouTube channels containing political/cultural content, which they organise into political categories. They find no evidence that viewers watching a video from a politically moderate channel are recommended videos from a more extreme channel: in fact, recommendations flow in the other direction. Again this argues against a role for recommender systems in political radicalisation.

But again, there are methodological problems with API studies. Social media APIs do not surface fine-grained information about recommendations: in particular, the information provided by YouTube's API is about recommendations made to an 'anonymous user', with no browsing history. But the whole point about a recommender system is that it makes different recommendations to different users, based on what it knows about their past behaviour on the platform. APIs that don't link recommendations to user histories simply don't allow us to study the effects on users that we are concerned about.

Robot user studies simulate individual users who have a tendency to follow recommended links. They ask: if users *do* follow the recommended links on a social media platform, where do they end up? For instance, Ribeiro *et al.* (2020) deploys robot users on YouTube. These

researchers also focussed on a set of political YouTube channels, which they classified on a left-right scale, culminating on the right with three progressively more extreme categories of right-wing content: 'intellectual dark web', 'alt-lite' and 'alt right'. They found (among other things) that bots starting in the first of these groups can readily access the second group, and bots starting in the second group sometimes reached the third group. This result suggests that recommender systems *could* have a role in radicalisation.

Robot user studies are another useful information source about recommender system effects. But again, they have methodological problems. By and large, the robot-link followers follow very simple rules. They aren't intended to be good simulations of what people actually do: they just show what would happen *if users followed the links they were recommended*. Robot users typically have very impoverished browsing histories, so what the recommender system knows about them is very limited. Again, it is hard to draw conclusions from these studies about the effects of recommender systems on real people.

Finally, **intervention studies** ask volunteer users to behave in certain directed ways in their consumption of social media. Wolfowicz *et al.* (2021) recruited young adults in East Jerusalem, a population considered at risk for Islamic radicalisation. The subjects were all new to Twitter, and were asked to set up a Twitter account and use it over a trial period of four months. Subjects were randomly assigned to two groups. A control group were asked to identify themselves and their existing social networks when joining, to prime Twitter's recommender algorithm for users to follow. A treatment group were asked to start their account 'from scratch', so the recommender algorithm had no prior information. At the end of the trial period, subjects' degree of radicalisation was measured. There were no overall effects of group on radicalisation—but there were interactions between intervention group and another measure taken by the experimenters, categorising the structure of a subject's social network as 'internally focussed' (inward-looking) or 'externally focussed' (outward-looking). For the group receiving normal recommendations, 'externally-focussed' users were more radicalised than 'internally focussed' users—while for the group using a 'weakened' recommender algorithm, it was the other way round. This result suggests that recommender algorithms can have effects on radicalisation—*in conjunction* with other variables.

Intervention studies are extremely useful in testing *causal hypotheses* about the effects of recommender algorithms on user attitudes. Many external studies of user effects report *correlations*: for instance, Boxell *et al.*'s (2017) population study effectively reports a low correlation between users' level of exposure to social media and their degree of political polarisation. But the hypothesis we really want to test is that social media recommender systems have a *causal role* in moving users towards extreme positions. To test this hypothesis, it is necessary to conduct a study that *manipulates* the experience of the recommender system for different groups of users in different ways, and then looks for *differences* across these groups.

In general, the need for intervention studies to test causal hypotheses is widely recognised: an elegant statement of this principle is given by Pearl (2009). Intervention normally happens in the form of a *randomised controlled trial*, where subjects are randomly placed into different groups, that are exposed to different experiences. Often there is a 'treatment group' that receive a certain intervention, and a 'control group', that do not receive it. This is the paradigm in drug trials, for instance, where the control group receive a placebo. Unfortunately, 'external' studies of social media platforms cannot readily manipulate users' experience of recommender systems. Wolfowicz *et al.*'s study attempts a manipulation—but as they note, the sample of users they recruit is quite small, and may be biased in various ways. It is also hard to ensure that users in the two groups always behave as they were instructed.

To summarise: 'external' methods for studying effects of recommender systems on users' political attitudes suffer from a number of serious methodological problems. Population studies have problems with confounding variables; browser logging studies have problems with

sampling bias; API studies are hampered by limitations in the information surfaced about recommendations; robot user studies show effects that may not generalise to human users. None of these methods test properly *causal* hypotheses about the effects of recommender algorithms on users, because they do not *intervene* in users' experience of recommender algorithms, by *manipulating* this experience in different randomly-selected groups. The external studies that do attempt to intervene in relevant user experiences—such as that of Wolfowicz *et al.* (2021)—suffer from other methodological problems: in particular, they have problems with sampling bias and sample size, and it is hard to ensure the intended interventions play out as intended, because they rely on the cooperation of subjects.

In the light of these issues, it is not surprising that external studies have reached a variety of different conclusions about the effects of recommender systems on users' attitudes towards extremist content. Some studies, like those of Ribeiro *et al.* and Wolfowicz *et al.*, find recommender systems do have effects on attitudes to, or availability of, extremist content. Other studies, like those of Boxell *et al.* and Ledwich and Zaitsev, find little or no evidence for such effects. Still other studies, like that of Flaxman *et al.*, find evidence of small effects. Some of this diversity likely reflects the fact that recommender systems very likely have different effects at different times and places, on different groups of people. But some of it may also be due to inherent *methodological shortcomings* of the techniques available to external researchers.

2.2 A survey of 'internal' methods for studying recommender system effects

The situation within social media companies is very different. Social media companies *deploy* recommender systems to their users: they are ideally placed to manipulate user experiences of recommender systems, and study the effects of these manipulations. Indeed, all the major social media companies *do* conduct experiments of exactly this kind. They do so in different ways, and for various purposes. Platform users are effectively subjects in these experiments. They are not widely aware of this, but permission to be used in such experiments is built into the terms of service for all the major social media companies in one way or another. [Reference] In this section, we will outline the methods that companies are known to use in their own internal studies of recommender system effects.

2.2.1 'A-B tests'

A simple and widespread form of experiment for companies is **A-B tests**, also called **randomised controlled trials** or **RCTs** (see e.g. Shani and Gunawardana, 2011). In an A-B test, some selected set of users on the platform is identified, and its users are divided randomly into a number of groups. Each group is given a different *version* of the platform's recommender system. There are many aspects of the recommender system that can be varied, so there is a large space of possible 'versions' to try. At the end of a trial period, in which users interact with their assigned version of the recommender algorithm, a set of measurements of user behaviour are made on all the users in the study, and averages over these measures are computed for each group. If there are significant differences between groups on any given measure, they can be reliably attributed to differences in the version of the recommender system they were exposed to.

A key point to make about A-B tests is that they *control* very effectively for the myriad factors that influence user behaviour alongside recommender systems. In particular, they control for the agency of 'political influencers', who operate on platforms to recruit followers, using a variety of strategies. Commentators often identify human influencers as a complicating factor in analyses of recommender system effects (see e.g. Llansó *et al.*, 2020). Influencers cer-

tainly exert important effects. But the groups of users in an A-B test are all equally available as potential targets to human influencers. If one recommender algorithm is easier for human influencers to exploit than another, this is still a fact about the recommender algorithm.

What do companies use A-B tests for?

A-B tests are used by companies to explore which versions of the recommender system are 'best', from some company-internal perspective. The criteria that companies are looking for are diverse and complex, but a central concern for companies is to find recommender systems that most effectively learn to give users *the content they want*. When we introduced a schematic recommender system in Section 1.2, we actually defined it in those terms, as a system that observes what kinds of content each user chooses to 'engage with', and then prioritises similar kinds of content in that user's feed.

The fundamental idea here is to find the best way to *keep users on the platform*. This is ultimately a commercial matter for companies: the longer a user spends on their platform, the more advertising they can sell to that user. A key role of A-B tests is to search for an *optimal* version of the recommender algorithm, that keeps users on the platform *most effectively*.

Again, exactly what is being optimised is a very complex matter, that differs from company to company, and is an important part of company IP. How 'engagement' is measured is also a complex matter. And user engagement is likely not the only measure that companies optimise for. Other factors include 'meaningful social interactions' and 'user satisfaction'; see Stray (2020) for a review. Our key point is just that A-B tests provide a very effective instrument for *comparing* different versions of a recommender algorithm, in terms of the effects they have on user behaviour. And in particular, they provide a much better way of studying recommender system effects than the 'external' methods we reviewed in Section 2.1.

Advantages of A-B test methods over external methods

A-B tests have several methodological advantages over the 'external' methods of studying user effects of recommender systems. We'll briefly review these here.

First, the study can be run on very large groups of users. The pool of potential users for an A-B test is as large as the platform's user base. In practice, small subsets of users are studied, but these user groups are often still far larger than the groups studied in external experiments.

Second, there are *no selection biases* to contend with in A-B trials. External experiments often rely on volunteers, who are a skewed sample of the user population as a whole: it's hard to know if the results of experiments on volunteers extend to the full user population. With company-internal A-B trials, on the other hand, all platform users can be equally easily included in any trial, because they have all given their consent by agreeing to the terms of service.

Third, there is every opportunity to *control for confounding variables* in the groups created in an A-B test. By assigning users to groups at random, and using large groups, experimenters can be relatively confident that the only systematic difference between groups is in the intervention they experience: so differences between groups can reliably be attributed to the intervention. Experimenters can also choose to control explicitly for a wide range of potentially confounding variables, because companies have a great deal of information about their users. (It is particularly important to control for time and place: there is very good evidence that the influences of social media are conditional on these factors, as discussed in Section 2.1.)

Fourth, experiments do not rely on subjects in different conditions doing as they are instructed. The manipulation that is applied happens within the social media platform; they just use the platform as normal, typically not even aware an experiment is being conducted.

Fifth, experiments done inside large companies can make use of rich, sophisticated measures of *user behaviour*. In the domain we are concerned with, experiments should study users' attitudes towards extremist content—and towards TVEC in particular. Users *express* these attitudes in obvservable behaviour towards content items on the platform (clicks, 'likes', shares, comments, and so on). Companies have an unrivalled view of this behaviour, compared with external researchers—even those using browser loggers or APIs. But in addition, large companies have rich resources for *classifying* content items—and in particular, for recognising items as 'harmful content' of various types.

Finally—and to reiterate—A-B studies explicitly *test causal hypotheses* about the effects of recommender systems on users. They manipulate recommender systems in different ways for different groups: any differences between user group in measures of user behaviour can be *causally attributed* to these different experiences, so the experiment provides direct information about the key question at hand: the *causal effects* of recommender systems on users.

We should stress that A-B studies have methodological problems of their own. In particular, users from different groups can communicate with one another, which would tend to blur any differences arising from different treatments. But there are also several ways of addressing this problem (see Eckles *et al.*, 2016 for a good introduction), and it is relatively minor compared to the problems faced by external methods.

2.2.2 Huszár *et al.*'s controlled study of recommender system effects

In most large companies, A-B tests are used to optimise the recommender system, as just described in Section 2.2.1. But this isn't always the case. In Twitter, for instance, the recommender system appears not to be experimentally optimised in this way. However, Twitter did conduct at least one experimental intervention of their own. In 2016, when they first introduced a recommender system that learns user preferences, they created a 'control group' of users who were not exposed to the recommender system. Instead, these users continued to see new content items in reverse chronological order. This control group have never been exposed to the recommender system. In 2021, Twitter released a study that compared this control group of users with a 'treatment' group of users who were exposed to the recommender system. The study was published in PNAS early this year, after peer review (Huszár *et al.*, 2022); we'll refer to this version of the study. The study was effectively a single, very large A-B test, where the intervention was either to provide, or withhold, the platform's recommender system. To give an idea of scale, over 11M users (1% of all Twitter users) were assigned to the control group, and over 46M (5% of all users) were assigned to the treatment group. A dataset was created logging all the tweets seen by each user in the study, during a time period in 2020.

We are particularly interested in Huszár *et al.'s* Twitter study, because its findings were *made public*, in a commendable initiative to provide transparency about recommender system effects. Huszár *et al.*'s study asked whether recommender systems *amplify political content* in users' feeds. They classified tweets into political categories, by two methods. First, they grouped the tweets of politicians according to their political party, in several selected countries. Second, they grouped tweets from all users that linked to articles originating from US media outlets, by assessments of the political leaning of these outlets. For all of these classes of tweets, they asked whether Twitter's recommender system 'amplified' tweets of this class—that is, whether users in the treatment group (exposed to the recommender sys-

tem) encountered more of these tweets than users in the control group (exposed to a reverse-chronological feed). Their basic finding was that the recommender system amplified *every* category of political content, across all countries they examined. The effects were startlingly large: users in the recommender system group were often more than twice as likely to see political items than users in the control group. Moreover, there was more amplification of right-wing politicaians' tweets than those of left-wing politicians in all but one country studied. (The focus was on Western countries and Japan; Germany was the exception in this case.) In the US dataset, there was more amplification of 'partisan' political content than of centrist political content—again, with particular amplification of partisan right-wing content.

We would like to commend Twitter for their decision to make the results of their study public. Huszár *et al.*'s study is extremely important, both as a demonstration of how company-internal methods can be productively and safely used to provide information to the public about recommender system effects, and in the first-order information it surfaces about recommender system effects. We will reflect on these contributions separately.

Huszár *et al.*'s study as a methodological benchmark

From the perspective of methodology, Huszár *et al.*'s study serves as a paradigm example of how company-internal methods can be used to provide high-quality information about the effects of recommender systems. Huszár *et al.*'s experimental method has all the advantages over external methods that we enumerated for A-B tests in Section 2.2.1: a huge sample size, unbiased selection of subjects, few confounding variables, rich behavioural measures, and a design that delivers findings about the causal effects of recommender systems. It provides high-quality data.

Equally importantly, Huszár *et al.*'s method surfaces this data *to the public*, in ways that safeguard the rights of users, and the intellectual property of the company. In any transparency exercise, in which a tech company provides information to the public about some aspect of its operation, it is vital it does not disclose personal information about users, or (for separate reasons) information about its own intellectual property. In the case of Huszár *et al.*'s experiment, there is absolutely no danger of disclosure of personal information about individual users, because data is aggregated over very large groups of users, and over very large sets of tweets. Neither is there any danger of disclosing company IP: the transparency in this case does not bear at all on the *workings* of Twitter's recommender algorithm, but rather on its *effects on users*. How these effects are achieved is not disclosed in any way—and arguably is of far less interest to external stakeholders. In short, Huszár *et al.*'s study also serves as an example of how information about recommender system effects can be safely *disclosed*. The methodology gathers high-quality experimental data, and *also* allows this data to be safely provided to the public.

In relation to safe disclosure, Huszár *et al.* note (p5) that their project was reviewed by Twitter's legal and privacy teams, who determined that 'additional notice and consent mechanisms were not required'. Interestingly, they state that the experimental intervention conducted by Twitter (their creation of the 'control' and 'treatment' groups) was not carried out 'for the purpose of research', but rather 'for the business purpose of improving the algorithm': it is for this reason, apparently, that further consent was not needed. In Twitter's case, the critical consent required from users seems to be about the *interventions* they may experience as participants in experiments, rather than the questions that may be subsequently asked about the effects of these interventions. We will take up this point in Section 3.2.2.

Huszár *et al.*'s study as a signal of the need for further work

Huszár *et al.*'s study also indicates very clearly that recommender systems have *large effects* in the domain of political content. This finding strongly signals a need for further studies of their effects in this domain. Huszár *et al.* note themselves that their method could readily be adapted to study the effect of recommender systems on 'manipulation, misinformation, hate speech and abusive content'. In the same way, it could readily be used to study effects on users' exposure to, or attitudes to, TVEC. We will take up this possibility in Section 2.3.

Huszár *et al.*'s finding that partisan political content is amplified more then centrist content is consistent with our concern that recommender algorithms that learn from user behaviour may lead users towards domains of higher political emotion, negative messages, and messages focussing on political out-groups (see Section 1.3). But Huszár *et al.* also examine whether 'far-right' and 'far-left' content is amplified (in their terms) by Twitter's recommender system—that is, whether users in their treatment group see more of this content than users in their control group. They find that in countries with enough far-left or far-right politicians to study, the amplification of tweets from these politicians is generally lower than that of tweets from moderate politicians. So the recommender system does not preferentially amplify extreme political content, by their measures. Note, however, that Huszár *et al.* do still find *some* amplification of this extreme content. This is still significant: if they were simply measuring the amount of extreme content seen by users in their control and recommender system groups, users in the latter group would still be seeing more extreme content. If there is more amplification of 'partisan' content than 'extreme' content, that may simply be because only a small minority of users move towards extreme content: the tweets these users consume will only have a small impact on the amplification levels recorded in Huszár *et al.*'s study. Other questions remain to be asked: for instance, what effect does the recommender system have on users who already consume a certain amount of extreme political content? We will take up these questions in Section 2.3.

2.2.3 'Offline' tests using causal models of recommender system effects

The A-B methods discussed in Section 2.2.1 allow companies to optimise their recommender system, by trialling alternative 'versions' of the system and picking the best (by some company-internal criterion). But as we noted, the space of possible versions of a recommender system is large. In fact it is vast: recommender systems are very large systems, with many variable parameters.[1] A-B tests can only try out a few versions at at time on users, so they provide a very inefficient method of searching for the optimal version.

Many companies now rely on a new generation of optimisation methods, that deploy recommender systems on *simulated* platform users rather than actual users. These methods involve the construction of a *general model* of how recommender systems affect the behaviour of users over extended periods of time (see Bouttou *et al.*, 2013 for the initial paper, and GPAI, 2021 for a short account). Very briefly: the general model in question is learned from data about how actual recommender systems influence the behaviour of actual users

[1] A word about 'parameters' here. Recommender systems often incorporate large neural networks, that learn by adjusting the weights of connections between their neuron-like units. The number of connections in these networks is typically in the billions of connections. The adjustable connections of a neural network are often referred to as its 'parameters'. But we discuss the 'parameters' that distinguish versions of a recommender system version being explored during A-B testing, are not talking about these parameters. We are talking about parameters that distinguish whole network designs: for instance, that vary the types of input it receives, or the number of units it has, and how they are connected. These are often referred to as the system's 'hyperparameters'. It is values of these hyperparameters that are explored when a recommender system is optimised. But we will continue to use the term 'parameter' in the current report.

on a given platform. But this data is gathered in a way that allows the model to provide information about how 'counterfactual' recommender systems, with different parameter settings, would affect the same group of users. Note that the models that are learned are explicitly *causal* models of effects on users, that reflect data about the effects of interventions in user experiences.

With models of this kind, it becomes feasible to explore the space of 'possible' recommender systems much more efficiently, because the learned model can evaluate the performance of any possible system 'offline', without any new trials on actual users. Learned models of this kind in fact enable a whole new paradigm for optimising recommender systems, using techniques called 'bandit menthods' from the AI field of reinforcement learning to guide a search through the space of possible systems (see again Bottou *et al.*, 2013). But for our purposes, the main point is that the learned models that underpin these 'bandit methods' provide another way companies can study the causal effects of their recommender systems on users. Importantly, studies with learned models can be run 'offline', without involve new interventions on actual users. In practice, companies tend to use a mixture of 'online' A-B tests and 'offline' bandit methods to optimise recommender systems: a similar combination of methods could be used to study effects of recommender systems in areas related to harmful content and TVEC.

2.3 GPAI's proposed fact-finding studies

In our report last year (Knott, 2021), we argued that by far the best methods available for studying the effects of social media recommender systems on users are the 'internal methods' that are used by companies: the methods that we just reviewed in Section 2.2. Since these methods are only available within companies, we proposed that external stakeholders should *collaborate* with companies, to investigate the effects of recommender systems on users' attitudes towards extremist content—with a specific focus on TVEC. We argued for studies investigating effects relating to extremism because of the concerns summarised in Sections 1.2 and 1.3. We suggested a focus on TVEC in particular because companies are already participating with external stakeholders in several initiatives centred on content of this type: we will discuss these initiatives in detail in Chapter 3.

In our earlier report, we referred to the collaborative studies we called for as **fact-finding studies**; we will continue to use that term in the current report. We made some specific suggestions about the form of the collaborative study: in this section we will elaborate on these suggestions, to reflect work that has happened in the interim (in particular, Huszár *et al.*'s study at Twitter, which was released after our previous report was completed), and discussions we have had with companies over the past year.

The study we envisaged involved a group of **external researchers** collaborating with a given **company** to design and conduct a fact-finding study. We envisaged three phases. In a **design phase**, company staff and external researchers (operating under a suitable non-disclosure agreement) negotiate the technical form of the study to be conducted, so that it addresses the key question at hand (about the effects of recommender systems on user attitudes towards TVEC), and so it is demonstrably *safe* for the company and its users. In an **implementation phase**, company staff and external researchers (operating under another, possibly different non-disclosure agreement) collaborate in the running of the study, and the gathering of results. In a **dissemination phase**, external researchers and company staff collaborate on the publication of a research report describing the study and its results. In this section, we will elaborate on each of these phases.

Before we do so, a couple of preliminaries. Firstly, we want to emphasise that our proposed fact-finding study is not a proposal about how *laws* should be drafted in the area of social media transparency. We certainly think our proposed fact-finding study might *inform* discus-

sions about legislation, or discussions about how powers granted under legislation already developed could be profitably used. The term 'fact-finding study' is intended to convey the contribution we have in mind: our aim is solely to *gather relevant information* in this domain. We will elaborate on links with legislative processes currently under way in Sections 3.5.1, 3.6 and 4.5.

On a related note, the fact-finding study we propose is emphatically not intended to offer suggestions about how recommender systems should be *changed*, or 'improved'. There are some very interesting projects exploring these questions—but our project focusses squarely on methods for asking whether there is a *problem* with a given recommender system on a given platform (at a given time and place), in the domain of user attitudes to TVEC. We don't want to suggest that questions about changes or improvements to recommender systems can only be asked if problems are found: but we do think it's helpful to have methods that focus on identifying problems—and certainly that is the focus of our project. We'll elaborate on that point in Section 3.5.1.

We now turn to the three phases of our proposed fact-finding study, as it would play out in a given company.

2.3.1 The design phase

The design phase of the study, as we envisage it, would involve closely related discussions with three key groups in the company.

One discussion would happen with *company engineers*. This would address the *feasibility* of the study: what is the exact research question to be asked? What form of company-internal study is best suited for asking this question? What data would be reported by the study? How would the publication reporting the study be structured? It would also address the *practicality* of the study—in particular, how costly would it be? How long would it take? And it would address the *running* of the study: who would conduct the study? And in particular, how would external researchers be involved in this process? Some involvement by external researchers is necessary, to ensure the study takes place as it was designed. In our previous report, we referred to the external researchers fulfilling this auditing role as 'embedded researchers'. In those terms, what access would embedded external researchers need to allow them to fulfil their auditing function?

Another discussion would happen with *company lawyers*. This would address the *safety* of the study. In relation to platform users—does it safeguard the rights of these users? In particular, does it preserve the privacy of the personal data of these users? And is it consistent with the terms of service users agreed to when joining the platform? In relation to the company's own interests—does it preserve the company's IP?

A third discussion would happen with *company management*. Management approval would likely be needed to *initiate* the collaborative fact-finding exercise as a whole. It would also be needed to *approve* the form of the study negotiated by the engineers and lawyers (including any costs borne by the company), and possibly also the form of the accompanying publication.

We envisage the discussions with company engineers and company lawyers would be intimately linked. The discussion with company engineers may have to suggest several possible methods, for separate vetting by company lawyers against user and company safety criteria. The form of the publication reporting the study and its results may also be an issue for company lawyers to consider. The access granted to the 'embedded' external researchers during the implementation phase would also be an important matter for discussion with company lawyers.

Note the discussions we have in mind here would happen with *individual companies*. The details of the technical methods used to conduct the study will necessarily vary from one company to another, so separate discussions are certainly needed here. However, we can certainly envisage a phase of negotiations with multiple companies, that takes place as a preamble to design discussions with individual companies, in which a common framework for these design discussions is laid down.

2.3.2 The implementation phase

In the implementation phase, company engineers and embedded researchers (playing agreed roles) collaborate in carrying out a fact-finding study, of the form agreed during the design phase. (Note: it will be important to *pre-register* the form of the study, and the hypothesis tested, in advance, so that there can be no question that the company is selectively reporting results. A clear distinction between a design phase and an implementation phase is useful in identifying the moment when pre-registration should happen.)

Possible forms for the fact-finding study

The form of the fact-finding study could certainly vary from company to company. In a company like Twitter, the study would likely have the same form as the study of Huszár *et al.*, because Twitter does not conduct fine-grained A-B tests of different versions of the recommender system. At issue here would be whether users in the 'treatment' group of Huszár *et al.* (whose feed is curated by Twitter's recommender system) have a different experience of TVEC, or TVEC-related content, than users in the 'control' group (whose feed arrives in reverse-chronological order).

In platforms like Facebook or YouTube,[2] the study would likely involve A-B tests, asking whether users exposed to *different versions* of the recommender system have different experiences of TVEC, or TVEC-related content. There are some grounds for thinking this may be the case: for instance, the causes for concern discussed in Sections 1.2 and 1.3, or various reports made by company employees, most recently Frances Haugen from Facebook (see e.g. her testimony to the US Congress and to the UK Parliament). But these analyses and reports are no substitute for actual data from company-internal studies. A fact-finding study using an A-B design would provide data that speaks to exactly this question.

Two kinds of 'offline' study could also be used in this context. Firstly, an offline study using a learned model of recommender system effects (see Section 2.2.3) could possibly be employed to help design a suitable A-B test—for instance, by identifying recommender systems that differ in relevant user attitudes. But this would require a model of user effects that extends to behaviours that diagnose attitudes towards TVEC: whether there are such models would be a matter for discussion with company engineers.

Secondly, note that our fact-finding study can also be conducted 'offline' in a second sense: it can examine on *stored datasets of user behaviours* gathered during an earlier A-B test. Huszár *et al.*'s study of Twitter users runs on a stored dataset of this kind. Our fact-finding study could certainly run on a stored dataset, rather than as a live experiment on platform users, if records are kept from a suitable A-B test conducted in the past. One benefit of examining logged data comes from the fact that the content we're interested in, TVEC is *removed* from platforms as soon as it is identified. This makes it hard (perhaps impossible) to study user interactions with TVEC in a live experiment. But logs of user behaviour may include their interactions with TVEC content that was only identified later. If TVEC is preserved

[2] We will refer to 'platforms' in these cases, rather than to the companies that own them (Meta and Google, respectively).

in logs of user interactions taken during A-B experiments, we can study the behaviours of interest in these logs.

Another point to note is that analyses could also be made of subgroups of users *within* the treatment groups of an intervention study. In particular, as noted in Section 2.2.2, subgroups of users with different platform behaviours could be identified in each treatment group at the start of the intervention, and the effect of the intervention could be studied separately for these different subgroups. This design would allow us to ask whether the effects of the recommender system intervention are *modulated* by types of user behaviour.

Possible behavioural metrics to assess users' attitudes towards TVEC

Whatever the experimental form of the fact-finding study, it will have to assess users' attitudes towards TVEC, by some metric that operates over their measurable behaviour on the platform. In our previous report (GPAI, 2021), we proposed a range of metrics that could be used. These proposals still stand, so we won't review them in any detail: they include 'end-point metrics' that measure engagement with (or searches for) actual TVEC on the platform, and 'pathway metrics', that measure engagement with content that is 'adjacent' to TVEC by some criterion—for instance, content identified by a platform as 'borderline TVEC', or 'hate speech'. Details of the proposed metrics, and their advantages and drawbacks, are given in Knott *et al.* (2021) Sections 5.4–5.7. 'Pathway metrics' provide another possible way to deal with the likely sparsity of actual TVEC on a platform.

The best metrics to deploy in a fact-finding study would certainly be a matter for discussion with companies. They have great expertise in this area: all the large social media companies have well-established methods for identifying TVEC and adjacent content such as hate speech; typically there are specialised teams of engineers who develop and deploy these methods. It is interesting to note that the tasks of recommender system evaluation/optimisation and content moderation are often the preserve of separate teams within companies; the fact-finding study we have in mind would involve interactions with both teams.

2.3.3 The dissemination phase

After the fact-finding study has been conducted, we envisage a final phase, where a paper is published describing the study and its results.

The paper would again be a collaboration between the company staff and external researchers who were involved in designing and running it. As already noted, the content of the paper would be something discussed in advance by the collaborating parties, during the 'design phase', so there is clarity about what it will disclose. But we suggest the *form* of the paper should be that of a standard scientific report, that presents a research question, a set of methods, and a set of experimental findings. Transparency processes in a case like this should involve standard paradigms for empirical science, so that the methods used and the results obtained can be critically assessed by the wider community of stakeholders.

2.4 Summary

In this chapter, we summarised and expanded our argument from the previous report: that by far the best way to study the effects of recommender systems on users is through the methods used by the companies who implement these systems. We presented an extended conception of a fact-finding study, whereby external stakeholders can collaborate with companies to surface information about recommender system effects. Our fact-finding study

focusses on the important domain of TVEC, which already brings together companies and external stakeholders in meaningful collaborations, that look beyond company-internal objectives, towards measures of social good.

The fact-finding study outlined in Section 2.3 has several things to recommend it. We'll conclude by recapping these.

Firstly, it uses *the best methods available*, to study a question of great social importance, that we do not yet know the answer to. The major social media companies have already indicated the importance of addressing this question.

Secondly, the technical methods it envisages are *already in widespread use*, by all the major social media companies, in one form or another. We do not propose the development of new methods: in terms of expertise, and resourcing, the proposed study would involve relatively small extensions to existing processes and implementations within companies.

Thirdly, these same technical methods also promise to provide ways of *safely disseminating* answers to the outstanding questions. As we noted in Section 2.2.2, these methods aggregate highly abstract datapoints over very large user groups; they also provide transparency about the *effects* of algorithms, rather than about their internal workings. Huszár *et al.*'s study of Twitter is eloquent evidence that studies of the kind we envisage do not disclose personal data about users, or company IP.

Of course, each company must do its own due diligence on the relevant legal issues (user rights, company IP, researcher access). But—as a final point—the discussion mechanisms we propose for the fact-finding study provide a *concrete focus* for these discussions. In the design process we envisage, the relevant legal discussions will be about fully specified study proposals, expressed with detailed reference to the company's existing mechanisms for evaluating recommender systems. This should be helpful in focussing the legal discussions.

3 Discussions about GPAI's fact-finding study, and related transparency initiatives

We have discussed our proposed fact-finding study with several groups over the past year. In this chapter, we'll report on these discussions, and also on other discussions taking place around the world about ways to provide transparency about social media recommender algorithms. Our aim is partly to report on the work we have done this year. But we also aim to *reflect* on the complex discussions that are currently under way around oversight of recommender systems. We want to offer participants in these discussions a holistic view of what is going on, and to offer some thoughts about productive directions, both for the conversation in general, and for GPAI in particular.

The discussions we will review have taken place in a very wide variety of settings. Some have involved particular pairs of research groups and companies. In Section 3.1 we'll describe an instance case of one such discussion, between our group and Twitter. Other discussions take place in large international groupings, that bring many governments, tech companies and citizens' groups to the same table. We have participated in two broader discussions of this kind: one organised by the Global Internet Forum to Counter Terrorism (GIFCT), and one organised by the Christchurch Call to eliminate TVEC online. We'll summarise discussions at GIFCT in Section 3.2. Discussions at GIFCT brought us into contact with several government groups that we have subsequently talked to individually: we'll summarise these discussions in Section 3.3. And we'll summarise our discussions in Christchurch Call workstreams in Section 3.4.

In Section 3.5, we will review some discussions that have taken place without our strong involvement, but that have addressed our fact-finding project proposal in one way or another. We conclude in Section 3.6 by discussing regulations being developed that include provisions for transparency around social media recommender systems, and in Section 3.6.6 by noting some voluntary codes of practice for companies that mention recommender systems.

3.1 Discussions between GPAI and Twitter

Our discussions with Twitter were brokered through the New Zealand government, which has a close relationship with Twitter, particularly since the terrible events of March 2019, when a terrorist livestreamed his attack on two mosques in Christchurch. Twitter has been very supportive of efforts to prevent the dissemination of TVEC. Our team has had several meetings with Twitter's Machine Learning, Ethics, Transparency and Accountability Team (META). The co-leads of our working group (GPAI's Responsible AI Working Group) also liaised with senior policy directors at Twitter. These interactions led to Twitter informally proposing to conduct our fact-finding study, using the same methodology as the study of Huszár *et al.* (2022) (see Section 2.2.2).

Twitter's informal proposal was that the running of our fact-finding study should be connected to another programme, developing new privacy-enhancing technologies (PETs) to enable external researchers to access the company's data. Twitter's PET project was announced in January this year: it involves a partnership with an open-source community called OpenMined, which develops a PET platform. Our initial discussion with Twitter about how to use OpenMined's methods to run our study didn't continue. However, a more recent and broader

initiative, the 'Christchurch Initiative on Algorithmic Outcomes', consolidates the idea of a partnership between Twitter and OpenMined to provide transparency about recommender system effects using PET methods, and includes additional partners. We'll describe this initiative in Section 3.4.2.

We are certainly interested in PET technologies as a framework for enabling collaborations between companies and external researchers, of the kind needed to conduct our proposed fact-finding study. We are very willing to collaborate in projects involving PET methods. GPAI has a great deal of expertise in PET: its working group on Data Governance is involved in a large project on PET; for details, see its 2021 report, and an announcement from June this year.

We do want to signal three points about PET use in the specific context of recommender system experiments, however. Firstly, the question of timing is a crucial one. How long will it take to develop the required PET platform? The need for information about recommender systems is very pressing, and delays should be minimised. Even small delays in running pilots could be significant, because pilots could usefully inform legislative processes that are currently under way, as we discuss in Sections 3.6 and 4.5. If a PET interface is still some way off, we think it would be valuable to conduct our study without it. (PET technology is clearly not essential in the study we have in mind: the study Huszár *et al.* conducted did not need it, and our proposed study envisages using exactly the same methods.)

Secondly, PET technologies still require some access to company systems by external researchers playing an auditing role. It may be that they helpfully structure the role of external auditors, so the kind of 'embedded researchers' we envisioned in Section 2.3.2 are not needed for every study. But some access by auditors will be an essential part of any mechanism providing transparency about recommender system effects.

Thirdly, we still foresee an important role for discussions with company engineers in broader discussions about recommender system transparency. It may well be that such discussions require some external researchers with some degree of knowledge of, or access to, company systems.

3.2 Discussions at GIFCT

GIFCT was originally founded as a partnership between tech companies, to facilitate the sharing of information about TVEC online. To further this mission, it convenes multi-stakeholder groups to work on key challenges at the intersection of terrorism and technology. As part of GIFCT's work to support its members in delivering on their commitments taken as part of the Christchurch Call community, it has also become a key forum for discussions about possible roles of social media recommender systems in the development of extremism.

This year, three of GIFCT's working groups considered issues relevant to recommender systems: the Transparency Working Group, the Legal Approaches Working Group, and the Technical Approaches Working Group (in which we participated, through our co-lead, Ali Knott). In this section, we'll review discussions and recommendations made by these three groups, presented in reports delivered at the GIFCT Summit in July this year. We'll also review a report written by a researcher participating in all of these working groups, released at the same event, that draws together the discussions of algorithmic amplification that arose across groups.

3.2.1 GIFCT's Transparency Working Group

GIFCT's Transparency Working Group ran a whole project on recommender systems. The report for this project (Whittaker, 2022) takes the form of a literature review of experimental work into effects of recommender system effects on user attitudes towards TVEC.

The review covers similar ground to the literature review we provided in our first report (GPAI, 2021 Chs 3 and 4). But it focusses tightly on effects related to extremism, in line with GIFCT's explicit focus on TVEC: effects of recommender systems in the area of political polarisation and misinformation are not in scope. It also describes some studies published after our review (and a few studies we missed).

Like our review, Whittaker *et al.*'s review focusses on methods. Their report notes many of the same methodological shortcomings as we noted in our report: in particular, there are few studies that involve controlled trials, and fewer still that involve experimental manipulation of recommender systems. Whittaker *et al.* concur with our argument that manipulations of a recommender system are necessary to properly study its causal effects on users. One of their key recommendations is for more collaborations between external researchers and companies, very much in line with our proposals. They also have other useful recommendations that go beyond the areas we considered: in particular, they recommend more research is done to assess the effectiveness of companies' changes to their recommender algorithms, and that more focus should be given to TVEC in languages other than English.

3.2.2 GIFCT's Technical Approaches Working Group

GIFCT's Technical Approaches Working Group[1] also conducted a dedicated project on recommender systems this year. Our GPAI group was involved in this project, through participation of our co-lead, Ali Knott. Ali also had input into the project's report (Thorley *et al.*, 2022), though he was not one of the penholders.

A focus in this project was on moving from *discussions* about possible transparency mechanisms to *trials* of possible mechanisms, running inside companies. Building on earlier literature reviews, in particular a review produced by a related GIFCT working group last year (CAPI, 2021), the project seeks to define some practical studies that companies could run, to trial methods for surfacing relevant information about recommender systems. The report refers to these trial studies as **pilot studies** of transparency mechanisms. It proposes three concrete pilot studies that companies could conduct—one of which is GPAI's proposed 'fact-finding study', as introduced in the current document in Section 2.3. In this section, we will outline each of the three proposed pilots, and the discussions that arose about them in the project. But before we do, we'll make a couple of preliminary notes.

Firstly, the concept of a pilot study is in itself something of considerable value. The term 'pilot study' clearly refers to an implemented exercise carried out by a company—but it also indicates clearly that the exercise is *experimental*, and serves to further a broader discussion about methods, rather than to provide definitive data in its own right. It is a useful way for companies to explore concrete transparency methods. Perhaps for this reason, the concept of 'pilot studies' of recommender systems in companies has spread beyond GIFCT discussions to many other fora in the past year. It surfaced in discussions between the UK and New Zealand (see Section 3.3), in the agenda for the Christchurch Call Summit (Section 3.4), and in the recent Christchurch Call Algorithms Partnership (Section 3.4.2).

Secondly, because the idea of pilots originated in a group *at GIFCT*, the group's discussions and recommendations about pilots come from an interesting group of stakeholders. Group

[1] The full title of this working group is 'Technical Approaches: Tooling, Algorithms and Artificial Intelligence'.

members included representatives from two of the big tech companies, from several governments, from citizens' groups, and from academia. This means that the project report (Thorley *et al.*, 2022) report provides an interesting preview of issues that are likely to arise in forthcoming discussions about pilot projects.

As already noted, one of the pilot projects proposed in Thorley *et al.*'s GIFCT project is GPAI's own 'fact-finding study'. We'll describe the proposed pilots out of sequence, starting with our own, which we can present rapidly.

Pilot 2: Effects of recommender algorithms on users

Thorley *et al.*'s 'Pilot 2' project is effectively the GPAI fact-finding study. The question it addresses is 'What are the effects of recommender systems on platform users' attitudes towards TVEC?' Two forms of the study are suggested, similar to the forms proposed in Section 2.3.2 of the current report. It is suggested that the study's results are disseminated in a scientific paper, as recommended in Section 2.3.3 of the current report.

Pilot 2 raises various questions about access to user data: on what basis are researchers allowed to access data about user behaviour, so as to evaluate effects of recommender algorithms on this behaviour? We will summarise discussions on this topic below, after introducing all three pilots.

Pilot 1: Effects of users on recommender algorithms

Thorley *et al.*'s 'Pilot 1' project asks a question that is in some way the converse of the Pilot 1 question: it aims to study the effects of *user behaviours* on *recommender algorithms*. Specifically, it asks 'What user behaviours are likely to prompt the recommender system to recommend *TVEC-related content* to them?'

As noted in our introduction to recommender systems in Section 1.2, influences between a user and a recommender system run in both directions. The recommender system influences the user's behaviour, by curating the user's feed; and the user influences the recommender system, by supplying evidence about the kind of content s/he likes to engage with, that periodically contributes to its training data. Thorley *et al.*'s Pilots 1 and 2 address these two separate influences.

Pilot 1 seeks to study the influence of user behaviours on recommender system offerings. The notional design involves the identification of several *groups* of platform users, that have different online behaviours. Different groups might consume content of different kinds, or subscribe to different channels. An initial option considered is to identify groups of *actual* platform users, using a browser-logging paradigm of the kind discussed in Section 2.2, and then to divide these users post-hoc into different groups, based on their behaviours. The recent study of Chen *et al.* (2022) is given as an example of this paradigm. Note that if this method was conducted internally to a company, it would raise the same questions about data access as arise for Pilot 2: on what basis are researchers allowed to access the user behaviour data needed to place them into groups? We will discuss this data access issue below. But identifying user groups post-hoc by their behaviour also raises a methodological concern: we'll focus on this concern to begin with.

If we are interested in studying *causal effects* of user behaviours on recommender system offerings, Thorley *et al.* argue we need to *intervene in user behaviours*, by the same reasoning that motivated the need for an intervention design to study the causal effects of recommender system design on user behaviours. The point stressed by Thorley *et al.* is that intervening in the behaviour of actual platform users is clearly not possible, either ethically or practically.

Considering both data access and methodology issues, Pilot 1 proposes to study *simulated users*, rather than actual users. The proposal is to create two or more groups of simulated platform users, with different platform behaviours, and observe what differences arise in the recommender system's offerings to these different groups. The aim is to discover what behaviours of a simulated user—if any—lead the system to recommend *borderline TVEC* to that user. The simulated users could be deployed on an actual social media platform, but they could also be deployed on a *simulated* social media platform. Some companies have elaborate simulations of their whole platform, including large interacting communities of simulated users, and a simulated recommender system. (For instance, Facebook's simulation of its platform is described here.) Running the study on a simulated social media system would keep simulated users away from real users.[2]

While Pilot 2 asks in general terms if the operation of a recommender system has any effect on users' attitudes towards TVEC, Pilot 1 aims to *localise* these effects (if they exist) in particular kinds of user behaviour. Both pilots surface potentially valuable information, and the information they would offer is complementary. Note there are also interesting ways the two pilot designs could be combined. In Pilot 2, user groups are exposed to different recommender system experiences on the platform: but in Section 2.3.2 we noted that analyses could also be conducted *within* these groups, focussing on users with particular behaviours. This builds in some aspect of Pilot 1's attention to user behaviours, and may allow us to see if effects of recommender systems are *modulated* by types of user behaviour. User behaviour in our proposed design would be identified post-hoc, of course, rather than created by intervention.

We'll conclude with a few thoughts about Pilot 1, that Thorley *et al.*'s report does not fully cover. The first is a question: why does a study of simulated users have to be conducted internally to a company? As we saw in Section 2.1, many studies of 'robot users' have been conducted by external researchers. Such studies certainly 'intervene' with recommender system inputs in planned ways, so they legitimately test causal hypotheses about effects of user behaviours on system recommendations, even though they are conducted externally. The reason for an 'internal' study of robot users must hinge on the ability of companies to create more realistic user simulations. But we're not sure how realism helps provide good answers here. The recommender system is an *algorithm*, that partitions users into different groups, and offers different content to different groups. Experiments varying user behaviour are essentially probing the recommender system to investigate the groups it has formed: the behaviours of simulated users only need to be good enough that the recommender system can identify one of the groups it has learned about.

Secondly, we're not sure that an intervention experiment is necessary to test causal effects of user behaviour *on recommender system behaviour*. Intervention is certainly necessary in our fact-finding study (Pilot 2), because user behaviours on a platform arise from a huge diversity of different causes in their online and offline lives, and our aim is to *isolate* effects due to one particular factor (their recommender system experience). Pilot 1 is interested in what causes a certain output to be produced *by an algorithm*. An algorithm's output doesn't have a multiplicity of unknown causes, that need to be controlled for. An algorithm's output is caused *by its input*, and nothing else: this is something we know from the nature of the system. If this line of reasoning is correct, then we can study effects of user behaviours on recommender algorithm offerings using post-hoc analyses of user behaviour. In this case there are clear advantages to a company-internal study: companies can study large groups of users, using detailed behavioural metrics. There are certainly questions to discuss about data access in such studies—the same questions that arise for our study (Pilot 2). We wonder

[2] If simulated users engage with other users on an actual platform, this would create problems of its own: simulated users of this kind would essentially be 'bots'. Social media platforms allow bots in certain specific contexts: there are some well-known uses on Facebook, for instance. But we can be sure that experimental bots enacting extremist behaviours would not be allowed on any of the major platforms.

whether post-hoc studies on real users, conducted internally to companies, might be a better way to study the causal effects of user behaviours on recommender algorithm outputs.

Pilot 3: An auditing study of of content moderation and content recommendation

Thorley *et al.*'s 'Pilot 3' asks about the effect of content moderation processes on recommender outputs. When a given content item is removed or flagged on the platform, what effect do these moderation actions have on recommender system offerings?

Pilot 3 proposes a form of auditing study to address this question, that essentially takes the form of an expanded 'transparency report' about TVEC and borderline TVEC on the platform, of the kind that companies already produce. The study presupposes an analysis of platform users into certain groups (not specified in the report), an analysis of content items by type (TVEC, borderline) and moderation action (flagging, removal, etc), and an analysis of engagement methods (viewing, sharing, liking, etc). It would report a *breakdown* of 'engagement events', by user group, content type, moderation action and engagement method. It would also report a similar breakdown of 'recommendation events', where the recommender offers an item to a user. Crucially, these analyses would also give separate breakdowns for engagement and recommendation events that occurred *before* and *after* the content item was moderated. The study thus broadly examines the effects of moderation on recommendation. Its key aim is to add some detail to existing transparency reports asserting 'we have reduced views of content items that were subsequently removed by 80%'.

Issues raised for discussion with Pilots 1–3

There was considerable discussion about all three pilots, both about privacy and consent issues (from companies and citizens' groups), and about issues of technical feasibility and resourcing (from companies). **Pilot 3** raised the most concerns overall. From a privacy perspective, there were worries that even though the data reported would be aggregated and anonymised, the amount of data surfaced, and its structured format may still pose risks to users' privacy: see Rocher *et al.* (2019) for a recent survey of relevant 're-identification' methods. From a technical perspective, the study would require data about user and recommender system actions to be stored in a revised format, which would require considerable effort to create.

Pilot 1, with a focus on simulated user studies, raised the fewest privacy concerns. Simulated user studies were recognised as posing no risk to the privacy of actual users. There were some technical concerns, however: after consultation with companies, no suitable 'synthetic environments' were identified that had 'sufficient access control for external researchers'. (Note the existence of simulated environments is not in question, at least for some platforms: in these cases, what is at isssue is how these environments can be accessed.)

Pilot 2, the GPAI fact-finding study, raised a mixture of privacy and technical concerns. The technical concerns were partly about the sparsity of TVEC content, due to the protocols in place for its removal. But it was recognised that offline studies on logged data could be used. There was also some recognition that 'pathway metrics' measuring TVEC-adjacent content may be feasible. In each case, there was an understanding that further discussion would be productive. There was also a technical question about how readily companies' A-B testing software platforms, designed for 'product improvement', could be adapted to study effects on an arbitrary user behaviour. But again, there was an understanding that further discussion would be productive.

There were also privacy and consent questions about Pilot 2. As framed, the pilot study would be conducted on users without 'additional consent' being sought: questions remain as to whether additional consent is needed. Again, it was acknowledged that further discussion of these questions would be productive. Some companies clearly allow Pilot 2 type studies to run without additional consent: Huszár *et al.*'s Twitter study is a case in point, as discussed in Section 2.2.2. It was recognised that the consent issues for studies like Pilot 2 are new ones, that require further discussion. The key novelty is that Pilot 2 expressly makes use of experimental interventions the company *already conducts* (or has already conducted), for its own purposes, and that users already give their consnt for under the terms of service.

There were also concerns about disclosure of private user data. But these were primarily about disclosure *to embedded external researchers*, rather than disclosure to the public. It was recognised that the format of published results posed little danger of disclosure of personal information.

In summary, it was recognised that the technical and privacy/consent issues that were raised about Pilot 2 could all be the subject of productive discussions between stakeholders represented in the group: companies, governments, researchers, and citizens' groups.

Report recommendations and discussion

Thorley *et al.*'s report made recommendations about each pilot, which we'll summarise here.

It was recommended that **Pilot 3** be 'rescoped and redesigned', and that groups be formed to carry out this work, by October 2022.

For **Pilot 1**, it was recommended that a research team be assembled, 'with the capacity to further the design and implementation of the project', by October 2022. It was also recommended that GIFCT should seek to arrange meetings between its research team and selected member companies, to explore the technical viability of this project.

Recommendations for **Pilot 2** were also for structured meetings with selected companies. It was recommended that GIFCT should seek to arrange meetings between its research team and specific member companies, 'to discuss technical aspects of the project', and separate meetings with the same member companies 'to discuss legal aspects of the project'.

We are waiting to hear from GIFCT about arrangements for these meetings: no meetings have yet been planned, that we are aware of. But the meetings will certainly be a productive step forward. For Pilot 2, in particular, meetings with engineers from selected companies, to discuss technical issues, would be very productive. There were no company engineers in our GIFCT group, so all technical concerns had to be relayed indirectly: a subsequent round of direct discussions would be far more efficient. Naturally, these discussions would be with engineers from a single company, and external participants would sign NDAs. Meetings with company lawyers would also be more efficient if they happened with individual companies, and focussed on a selected pilot project.

3.2.3 GIFCT's Legal Approaches Working Group

GIFCT's Legal Frameworks Working Group aimed to discuss general legal issues that arise for companies faced with requests for data access by researchers or civil society groups. A key structuring insight from their report was that there is no 'one size fits all' answer to questions of data access: the report accordingly focussed on access for one specific purpose, namely 'open-source' investigations of human rights abuses, associated with groups like Bellingcat. Discussions of legal issues were easier when focussed on concrete cases:

open-source investigations of abuses in Syria and Ukraine were used as case studies.

The group did not discuss legal issues for recommender system transparency. But we suggest that legal discussion in this area might also be facilitated if it could focus on concrete transparency scenarios, like the pilot projects proposed in the Technical Approaches Working Group, just discussed in Section 3.2.2.

3.2.4 Jazz Rowa's multi-group report on algorithmic amplification

GIFCT commissioned a security and governance expert, Dr Jazz Rowa, to participate in the three working groups whose findings we just described, to look at common themes relating to algorithmic amplification through the lens of human security. Her report (Rowa, 2022) examines the role of social media algorithms in radicalisation processes understood more generally, that take place in people and places and communities. The report reiterates some important themes from broader political discussions: there is no agreement about what counts as 'terrorist content'; current legal frameworks need updating to deal with current Internet technologies; users of these technologies also lack understanding of them. But its main contribution is in arguing persuasively that there is no clear distinction between 'online' and 'offline' spaces; and that the role of algorithms in radicalisation cannot be separated from other causal mechanisms operating in the world.

We fully agree that a full picture of radicalisation involves simultaneous attention to online and offline processes. But we also believe that there are good scientific methods for studying certain certain *components* of this complex system in isolation, if we can make use of company-internal methods. Our proposed pilot study does exactly this: it examines the effect on user behaviour of manipulating one isolated component of users' experience—the recommender system—while *controlling* for the myriad of other causal mechanisms that influence users' behaviour, both online and offline. The results that are surfaced by such an experiment admittedly only tell us about one tiny component of the wider radicalisation process. But they do, we believe, effectively *isolate* this component from the complex other factors that influence radicalisation. And they do so in a helpful way—because they identify what is in companies' power to *change* about the wider process.

3.2.5 Next year's GIFCT working groups

There is no dedicated working group on algorithms at GIFCT next year: but there is a working group on Frameworks for Meaningful Transparency, which aims to build on last year's work in this area. The work to be done will include a specification of how to interpret 'meaningful transparency' in the sphere of TVEC for different stakeholders. There are also working groups on the risks and safety-by-design best practices associated with various technologies, as well as on the best practices for designing positive interventions on a range of platforms. These working groups will each touch on various aspects of recommender systems.

The work to be done in the Meaningful Transparency working group will include 'a review of third party oversight models' and 'human rights considerations'. The third party oversight models will presumably include OpenMined's privacy-enhancing technology stack.

GIFCT is also working closely with the Global Network on Extremism and Technology (GNET), its academic research arm, to assess the user journeys in and around violent extremism on a range of platforms. This work will include an assessment of the role played by interactions with recommender systems in these user journeys.

3.3 Discussions with government groups

Our project liaises closely with the New Zealand government, which is taking a global lead on many issues relating to online proliferation of TVEC. Our fact-finding project as it was introduced last year (GPAI, 2021) was in fact framed as a case study of New Zealand users. This year we have extended the scope of the project to encompass more countries. On the technical front, company-internal studies can readily be deployed over users in many countries; again, Huszár *et al.*'s study is a case in point. On the political front, several countries are developing legislation around social media recommender algorithms (see Section 3.6); these countries are often particularly active in the discussions we have been involved in. We have had several discussions with government groups from particular countries, discussing potential collaborations. We have talked with a group in the French government, working on an API-based method for detecting political polarisation on social media, and a team at Public Safety Canada studying online radicalisation. We have participated in discussions at the Brookings Institute focussed on the US policy context. We are also interacting with an EU initiative, which we will discuss in Section 3.5.4. And we have had briefer consultations with groups in the Japanese and Italian governments. However, our most productive collaboration outside of New Zealand has been with the Online Policy Unit in the UK Home Office.

The Online Policy Unit, which is part of the Home Office's Homeland Security group, participated with us in GIFCT's Technical Approaches Working Group, and contributed one of the proposed pilots. To help maintain a focus on recommender system transparency, and on pilot studies in particular, as the GIFCT project drew to a close, we talked with colleagues in the UK and New Zealand governments who were negotiating the agenda for the meeting between the UK and New Zealand Prime Ministers in London that took place in July. The meeting did indeed cover these topics: and the joint statement released afterwards made a commitment to 'review the operation of algorithms and other processes that may drive users towards and/or amplify terrorist content', by means that include 'enhancing the evidence base through the delivery of research pilots'. Importantly, the joint statement also signalled that 'new measures' in this area would be announced at the Christchurch Call Summit in September.

The commitments made in the meeting between the UK and New Zealand Prime Ministers meeting created a useful focus for government groups in the leadup to the Christchurch Call Summit. We'll now turn our attention to the discussions on social media algorithms that happened within under the umbrella of the Christchurch Call this year.

3.4 Discussions in the Christchurch Call's Algorithms workstream

The Christchurch Call to eliminate TVEC Online was initiated in May 2019, by Prime Ministers Jacinda Ardern and Francois Macron, two months after the horrific attacks on worshippers at two mosques in Christchurch. It now brings together 58 governments and 12 tech companies, as well as a large advisory network of civil society organisations and academic groups. Members of our project also participate in this advisory network.

Companies and governments make several commitments under the call. Some are specific to companies: in particular, companies commit to 'review the operation of algorithms and other processes that may drive users towards and/or amplify TVEC'. Some are specific to governments: notably, to 'consider appropriate action' to prevent the online dissemination of TVEC. Some are joint commitments: notably, to 'develop effective interventions, based on *trusted information sharing about the effects of algorithmic and other processes* [our empha-

sis], to redirect users from TVEC').

The Christchurch Call runs several 'workstreams' to initiate and coordinate work relating to these commitments. Workstreams are defined at Summits, where country leaders and company executives meet to review ongoing work and create new workstreams. The key workstream for our project is the Algorithms and Positive Interventions workstream, which was initiated at the 2021 Summit. The objectives for this workstream, which we helped to define, include discussions about 'collaborative methods' for studying recommender effects in the domain of TVEC:

> We will design a multi-stakeholder process to establish what methods can safely be used, and what information is needed - without compromising trade secrets or the effectiveness of Online Service Providers' practises through unnecessary disclosure—to allow stakeholders to better understand the outcomes of algorithmic processes, and their potential to amplify terrorist and violent extremist content.

Several groups within the 'Call community' have prominently called for a process of this kind. Along with our GPAI project, Tech Against Terrorism have advocated for further scrutiny of recommender systems; the Institute for Strategic Dialogue (ISD) have pointed out the dangers of 'echo chambers'; the Center for Democracy and Technology recently announced project of its own on recommender algorithms.

3.4.1 The 2022 Christchurch Call Summit

The Call's 2022 Summit happened in September in New York: many of our project members participated online.

In the leadup to the Summit, meetings were organised with the Call Community to create the agenda for the Summit, along with discussion questions for leaders. We also participated in these meetings. There was a strong consensus in these preparatory discussions that the topic of recommender system transparency should be prominent at the Summit. Accordingly, one of the three Summit sessions was dedicated to this topic, and leaders' briefing notes defined some specific questions about recommender system transparency mechanisms for discussion.

The Summit did indeed make some concrete progress on this topic. As outlined in the joint statement released by Prime Ministers Macron and Ardern after the Summit, participants endorsed a commitment to 'develop shared solutions to studying algorithmic impacts within the Christchurch Call'—and, notably:

> to 'drive forward discussions on targeted research pilots that respond to questions raised by our Community, including on (…) the possible unintended consequences of human / machine learning AI interactions'.

Again, the concept of 'pilot projects' on recommender systems was clearly found useful in discussing and expressing new commitments made under the Call. Even more concretely, a specific new initiative on recommender system transparency was announced at the Summit by Jacinda Ardern, under the name of the Christchurch Call Initiative on Algorithmic Outcomes. We will describe this initiative in the next section.

3.4.2 The Christchurch Call Initiative on Algorithmic Outcomes

The Christchurch Call Initiative on Algorithmic Outcomes is a partnership between New Zealand, the US, Twitter and Microsoft, that focusses on particular methods for enabling external researchers to collaborate with companies to study recommender system effects. The specific aim of the initiative is 'to develop new software tools that will help facilitate more independent research on the impacts of user interactions with algorithmic systems'. The tools in question are the privacy-enhancing technologies (PETs) developed by OpenMined that we mentioned in Section 3.1: OpenMined is also a partner in the Initiative.[3]

As noted in Section 3.1, Twitter have been working with OpenMined since January this year. When our GPAI group discussed a pilot project on recommender system effects with Twitter, they proposed we conduct this pilot as a trial of OpenMined's PET software. So the basic plan to do this work has been in place for a while. The Initiative does nonetheless mark some important new developments. Firstly, the US government is now involved. This gives the Initiative considerable new prominence. Secondly, Microsoft is involved. It is interesting to reflect on Microsoft's interest in this project. Microsoft is certainly keen to study the role of recommender systems in the proliferation of TVEC: the blog post that announces their participation is in fact more explicit about the focus on recommender algorithms than the official Initiative statement. However, Microsoft is likely also interested in trialling the use of PET software in other domains. For instance, the Microsoft blog post mentions they envisage using it on their Azure and LinkedIn platforms to support *explanations* of *AI decisions*. This functionality appears to be directed at platform users, rather than researcher studying platform impacts.

We reiterate that we are looking forward to explore PET technologies as a framework for enabling external researchers to conduct studies of recommender system effects on users. But as noted in Section 3.1, we think discussions between companies and external researchers about recommender system studies should not be deferred until a PET platform is in place: they should proceed *in parallel* with PET software development. (In particular, the discussions about pilot studies that were recommended by GIFCT, as discussed in Section 3.2.2, should emphatically still take place.) PET technology is not necessary to conduct a study in the form we envisage, as Huszár *et al.*'s (2022) study demonstrates. And there is a pressing need for pilot studies to be conducted as soon as possible, to inform ongoing discussions about regulation.

It is also worth noting that conversations with company engineers are of great importance in designing good studies of recommender system effects, even if external researchers use PET to access their data. Conversations with engineers are of particular importance in making good decisions about the best behavioural measures to use in a given experiment. For instance, in studies about effects on user attitudes towards TVEC, input from experts in companies' content moderation teams, responsible for the classifiers that detect TVEC-related content, would be invaluable, perhaps essential.

3.4.3 Researcher data access initiatives

OpenMined is developing one mechanism for surfacing company data for external scrutiny. But there are many other initiatives in this area too. In this section, we will briefly review the initiatives that are most visible in current discussions.

[3] Elon Musk's acquisition of Twitter was announced just as this report was going to press, which creates some uncertainty about how this initiative will play out. In the current section, we assume it will continue as announced at the Summit.

OpenMined's privacy-enhancing technology (PET) is widely discussed at present. It seeks to enable digital platforms (including social media) to provide researchers enhanced access to data and analytics, while ensuring secure protection of company IP and user privacy. Traditionally this has been done through APIs and offline curated datasets. Offline datasets, being by nature retrospective and static, can rarely provide the dynamic experimental environment to replicate a real time online scenario of complex interactions on a digital platform. On the other hand, though APIs were designed to facilitate controlled query interfaces, they can be quite restrictive in their scope in order to maintain the privacy standards. The new PET technologies utilise the state-of-the-art innovations in safe and secure AI to provide access to real time and real life data, while maintaining cutting edge privacy standards.

Though the exact specifics of the technology components are still emerging, according to the OpenMined blog, the key elements of PET include differential privacy (Dwork and Roth, 2014), syntactic anonymisation techniques like k-anonymity (see e.g. Kenig and Tassa, 2012), homomorphic encryption (Mahato and Chakraborty, 2021), trusted execution environments (see Geppert *et al.*, 2022 for a recent review), secure multiparty computation (see e.g. Zhao *et al.*, 2019), zero-knowledge proofs (see Morais *et al.*, 2019), secure aggregation (Bonawitz *et al.*, 2017), federated learning (Kovnevcny *et al.*, 2016) and a set of conventional de-identification approaches such as masking, rounding, and hashing. These methods are incorporated into an integrated software platform called PySyft. This platform is open-source: OpenMined has created a public repository for this platform. PySyft decouples data privacy and data querying, using the methods enumerated above. Within well-defined limits set by the data owner, it allows an external party to ask queries about a dataset, and get meaningful answers, without getting a copy of the data itself.

OpenMined is looking to collaborate with leading players across different fields like healthcare and social media, and one of their industry partners is Twitter. According to the Twitter blog, they want to adopt PETs to (1) enable external researchers to access non-public Twitter data, and (2) internally democratise and scale their Responsible Machine Learning Workbench (a series of custom-built Machine Learning fairness and ethics tools used at Twitter). The Twitter META (Machine Learning Ethics, Transparency, and Accountability) team is leading this endeavour.

Differential privacy and Social Science One

The European Digital Media Observatory (EDMO) also wrote a extensive report on Platform-to-Researcher Data Access earlier this year, that follows the Code of Conduct defined under Article 40 of the European GDPR. The report emphasises the urgent need for safe and secure access to data by researchers as well as the broader civil society for social good. One of the major technical prospects of achieving that has been Differential Privacy, as put forward by Harvard's Social Science One initiative. Social Science One was established by Facebook in 2001 as a consortium of social science research institutes to study and reduce barriers to industry-academic partnership. Harvard University's privacy tools project webpage defines Differential Privacy as follows.

"*An algorithm is said to be differentially private if by looking at the output, one cannot tell whether any individual's data was included in the original dataset or not. In other words, the guarantee of a differentially private algorithm is that its behavior hardly changes when a single individual joins or leaves the dataset – anything the algorithm might output on a database containing some individual's information is almost as likely to have come from a database without that individual's information. Most notably, this guarantee holds for any individual and any dataset. Therefore, regardless of how eccentric any single individual's details are, and regardless of the details of anyone else in the database, the guarantee of differential privacy still holds. This gives a formal guarantee that individual-level information about participants in the database is not leaked.*"

This definition of Differential Privacy emerged out of a series of papers starting with Dinur and Nissim (2004) and culminating with Dwork *et al.* (2006).

3.5 A survey of other work on recommender system transparency and functionality

So far in this chapter, we have described initiatives we have participated in. In this section, we will introduce some initiatives on recommender systems that have taken place without our strong involvement, but which make some reference to our GPAI project.

3.5.1 The Toronto/Berkeley Facebook Newsfeed project

The Toronto/Berkeley Facebook Newsfeed project is a collaboration between academic researchers and Meta, to explore alternative ways of optimising Facebook's recommender system. The project is co-led by Jonathan Stray at Berkeley's Center for Human-Compatible AI and Gillian Hadfield at the University of Toronto's Schwartz Reisman Institute and the Vector Institute.

The project aims to develop a method for optimising Facebook's newsfeed recommender algorithm for some measure other than user engagement. The initial experiment, which is still under way, is trialling the method's effectiveness in optimising for a measure of 'online social support', which has been independently validated (see Nick *et al.*, 2018), and is plausibly something that changes to the newsfeed algorithm might affect. But the focus of the experiment is on the optimisation method, rather than this particular measure.

This project is the only study we are aware of in which academic researchers are working inside a social media platform to experimentally intervene in its recommender system. It is likely to produce very valuable findings about new ways for optimising recommender algorithms. But independently of any results the project delivers on this question, it is enormously valuable as an experiment into how external researchers can collaborate with companies. A complex mixture of technical and legal issues have to be addressed and worked through; new processes and new methods must potentially be defined. The study involves a diverse team of researchers, from academia, industry and civil society, with a mixture of technical and legal expertise. It is very much a 'pilot study', in the spirit of those proposed in GIFCT, that were discussed in Section 3.2.2.

The study takes place in the context of broader collaborations between academic researchers and Silicon Valley companies in the area of recommender system design. These collaborations recently resulted in a paper reviewing selected design methods and discussing issues that arise with these, both on technical and policymaking planes (Stray *et al.*, 2022). The writers of this paper only partially overlap with the researchers conducting the Toronto/Berkeley Facebook experiment. But they are similarly interdisciplinary: authors of the paper include scientists at Meta and Google, and academic experts from the fields of AI, technology law and mental health. And some of the design principles proposed in the paper fed into the Facebook experiment.

In particular, Stray *et al.* (2022) are keen to frame the *scope* of collaborative research in recommender system design so that it encompasses studies that explore potential benefits of these systems as well as studies that explore potential harms. As they say, 'focusing solely on harms is overly narrow'. They prefer to frame the general research objective as being to design recommender systems whose effects on users align with the *values* of those users (and their wider communities). Framing the objective in terms of human values allows for studies about how harmful effects can be avoided—but also for studies of how good effects

can be achieved. In this sense, this framing 'opens up new avenues for thinking about the role and responsibilities of recommenders in society'. The Toronto/Berkeley study is certainly exploring these broader avenues.

Stray and colleagues also have certain more methodological concerns about studies that focus on particular harms. Benefits and harms of recommender systems trade off in complex ways, so focussing on particular measures of harm arguably fails to present the full picture. In addition, it's hard to identify what harms are due 'to the recommender system', because there's often no obvious baseline to compare to.

On all these grounds, this is a good moment to re-examine our proposed fact-finding study. Our proposed study certainly has a narrower scope than Stray *et al.*'s (2022) research programme. It aims to study companies' *current practice*, rather than exploring how this practice could potentially be changed. It also focusses on measuring specific potential harmful effects on users. We think there's an important role for both types of study. But it's useful at this point to expand a little on the role we foresee for the narrower form of study we have in mind, that focusses on particular harms. We'll do that in the remainder of this section.

The role of a study focussing on possible current harms

As we see it, there are two important roles for studies focussing on possible harms caused by current recommender systems.

Firstly, measurement of harm is an important element of *reporting*, and accountability. The core motivation for our study is that companies should provide *more information* to external stakeholders about the effects of recommender systems. Any reporting protocol must name specific measures to be reported; protocols for reporting recommender system effects will be no different. And for recommender systems, protocols reporting measures of harm are particularly important. There are *prima facie* causes for concern about recommender system effects, that are supported by (admittedly imperfect) experimental evidence. In addition, there is considerable public and political interest in whether recommender systems have harmful effects, particularly in the area of TVEC, as just discussed in Section 3.4. The fact-finding study we are proposing is designed to provide a method for reporting good information on this question.

We want to be clear that we're not advocating laws that mandate companies to *avoid* harmful effects of recommender systems, or even minimise their harmful effects. There are some excellent analyses of why those laws would be hard, or perhaps impossible to draft. Keller (2021), comments on the difficulty of drafting laws that prohibit companies from 'amplifying harmful content', or that seek to assign legal responsibility to companies for the content they 'amplify'. Llansó *et al.* (2020) criticise regulatory proposals defining specific effects that recommender systems should achieve, or avoid. Stray's group also criticise such proposals, in particular in Thorburn *et al.* (2022). We'll discuss these criticisms more below.

But these difficulties don't make it wrong to develop ways to publicly study possible harmful effects of recommender systems. If a company's product is identified as causing a specific harm of some kind, there is an onus on that company (and other stakeholders) to look for remedies, even in cases that are known to be complex. Our key point is that if a harm is identified, then the question of how recommender systems should be *modified* then arises with a very specific focus. A second important role for the kind of study we have in mind is in providing *direction* to studies that seek to alter a company's recommender system.

Note that the methods being trialled in the Toronto/Berkeley experiment could certainly be deployed to minimise an identified harm, as well as to maximise a positively defined outcome. As we'll discuss below, positive and negative effects interact in complex ways, so the differ-

ence between studies seeking to minimise negative effects and studies seeking to maximise positive ones is not clearcut. But we feel it's a meaningful distinction nonetheless, especially given the importance of measuring harm in transparency protocols. In the wider field of AI ethics, there's certainly a well-understood distinction between research into 'AI for good' and research that identifies and attempts to fix problems that arise with AI. For instance, Floridi *et al.* (2018) distinguish a principle of 'beneficence' (to positively contribute to human well-being), and a principle of 'non-maleficence' (to avoid harms): they interact in complex ways, but are still worth distinguishing.

The role of focussing on specific harms

Any given recommender system has a complex mixture of positive and negative effects on its users. Stray *et al.* (2022) emphasise that studies of these effects should look at how they trade off against one another as the system changes, rather than focussing on a single effect, or a single harm. Our proposed study, by contrast, focusses on a specific measure of harm.

We agree that the debate should ultimately be about how the many different effects of a recommender system trade off against each other. But we also want to emphasise that this debate should be at least in part a *public* debate, with participation by external stakeholders, rather than one that happens within companies, out of sight. At present, the debate cannot even start, because we don't yet have good public data about *any* effects. One key objective in surfacing data about a particular measure of harm (such as user attitudes to TVEC) is simply to *initiate* a well-informed public debate. We see the role of our study as in *starting* a public debate—certainly not in concluding it.

We want to emphasise one point here. At present, the public don't know if modifications to a recommender system have *any* effect on user attitudes towards TVEC. This information simply isn't in the public domain. One piece of information we are critically missing is whether companies have *any agency at all* over user attitudes towards TVEC. If it were found that different recommender system versions cause appreciably different attitudes to TVEC, this would throw a whole new emphasis on companies' choices about recommender systems, and open up new areas for public discussion. And in due course, the fact-finding methods we advocate would help inform this discussion.

Issues with finding a meaningful 'baseline'

Stray and colleagues are also concerned by the perception in some quarters that the harmful effects of recommender algorithms could be readily eliminated by 'turning them off', and reverting to some neutral mode of operation. There are some study designs which allow some approximation of this. For instance, Huszár *et al.*'s (2022) study of Twitter compares a 'treatment' group of recommender-system users with a 'control' group of users receiving a reverse-chronological feed. But in many other cases, there is no natural control or 'baseline' condition against which to compare: designers have no choice but to compare the effects of different recommender algorithms *against one another*. For instance, a reverse-chronological content feed would make no sense at all on YouTube or Spotify. Even in the Twitter study, the 'control' condition has harms of its own, that need to be considered. These points are well made in an article by Thorburn *et al.* (2022). Similar points are made by Llansó *et al.* (2020) and Keller (2021), arguing against legislation prohibiting algorithms that 'amplify' certain content, because there is often no natural alternative to revert to.

We certainly don't think our fact-finding study would offer obvious solutions to any harm-related effects it uncovers. All it would do is to *report* on certain effects—and potentially to identify whether companies have any *agency* over these effects (and if so, to quantify the extent of this agency). Again, we see the study as initiating a public discussion about

recommender system effects, and as providing methods for furthering this public discussion.

Of course, the lack of clear baselines is common in public policy discussions. In discussions of road safety, there's no baseline speed limit policy, against which other speed limits can be assessed. There are just tradeoffs to be made among the benefits and harms of different speed limits. In their comments on baselines, Stray *et al.* are not criticising recommender system studies for not having clear baselines: they are criticising commentators who *expect to find* baseline conditions in these studies. So we have no disagreement with Stray *et al.* on this point. But to remain with the road safety analogy: policymakers deciding on speed limits do *need information* to make informed decisions about speed limits. They should have road accident statistics of various kinds. If such statistics were not available, there would be good reason to call for them to be produced. What we are arguing for in the social media space is *more quantitative information* about the harms caused by different recommender system experiences, and how these trade off against other effects, so a proper public dialogue can begin about these tradeoffs—and in due course, good policies.

3.5.2 The Action Coalition on Meaningful Transparency

The Action Coalition on Meaningful Transparency is an initiative coordinated by the Danish Government's Tech for Democracy Initiative. Its aim is to map the landscape of transparency initiatives currently under way, and to build connections between these. In due course, it will also make recommendations for improving transparency processes.

The initiative will bring together the full range of stakeholders, from academia, civil society groups, companies, governments and international organisations. Its structure provides for 'participants', an advisory council and a steering group: GPAI has participation in the steering group.

An initial focus of work will be to clarify definitions and terminology in the sphere of digital platform transparency, and to identify gaps in the work that is being done globally.

3.5.3 The Global Network Initiative's transparency project

The Global Network Initiative (GNI) is also working on transparency mechanisms. This organisation is primarily a grouping of tech companies, with a mission to 'advance freedom of expression and privacy rights in the ICT industry'. It also has membership from academia, civil society groups, and tech investors. (Crucially, governments are not members.)

An initiative on meaningful transparency is under way at GNI. In this initiative, discussions about transparency amongst GNI members will feed into discussions in the Action Coalition on Meaningful Transparency just discussed in Section 3.5.2. The relationship between the Action Coalition and the GNI is clearly a close one: the GNI is also a steering group member of the Action Coalition, and hosts the Coalition's website.

We're aware that GPAI's fact-finding exercise has been a topic in GNI's own internal transparency discussions. But we don't have any details.

3.5.4 Work in the EU Internet Forum

The EU Internet Forum was established in 2015, to address the misuse of the Internet for terrorist purposes. Amplification of extremist content is one interest of the group: a productive workshop on 'potential risks associated with algorithmic amplification techniques' was held in September last year, and a project on algorithmic amplification is currently under way.

Our GPAI group have had several discussions with researchers on this project, and will be formally consulted later this year.

3.5.5 The Council for Responsible Social Media

A recently formed group is the Council for Responsible Social Media. This group is mainly based in the US; it features prominent ex-officials from the US government, along with independent researchers and public health advocates. A key spokesperson is Frances Haugen.

The group's purpose is to 'advocate for key policies and legislation with the Biden Administration, on Capitol Hill, and in select state legislatures' in the US, and to 'publicly pressure social media platforms to make meaningful platform and internal governance changes'. These activities, of advocacy to governments and companies, are similar to the activities of our GPAI project, so we have something in common: we intend to make contact with them soon, to introduce ourselves.

Again like us, the group places social media transparency mechanisms centre stage. An immediate goal in its government advocacy is to push for Congress to adopt the Platform Accountability and Transparency Act, which we'll describe in more detail in Section 3.6.4.

3.6 A survey of regulatory initiatives involving recommender systems

In this survey, we briefly note a few jurisdictions where regulation is being developed (or is in place) that bears specifically on social media recommender algorithms. This is not a comprehensive review: we limit ourselves to the regulatory initiatives that have received most attention in the discussions we have taken part in. The EU's Digital Services Act is certainly the main focus for current discussions, but there are some other initiatives also worth mentioning. For a broader perspective on regulatory initiatives for social media bearing on TVEC and its removal, a useful resource is the OECD's recent review of Transparency Reporting on TVEC Online (see Section 4).

3.6.1 The EU's Digital Services Act

The EU's Digital Services Act (DSA) was proposed in December 2020, and was agreed on by the EU Parliament and Member States in April this year. It will come into force in January 2024.

The DSA imposes special rules on 'very large digital platforms', several of which relate to recommender systems. We'll summarise these rules, using the term 'companies' to refer to very large platforms. These are widely understood as imposing the strongest duties on companies in relation to recommender systems.

In the Recital of the act, introducing clauses that contribute to its interpretation, while not legally binding in themselves:

Recital (58) imposes on companies a duty to perform 'risk mitigation', that's specific to the dissemination of 'illegal content' (which inlcudes TVEC, by most definitions). This duty requires companies to 'consider (…) enhancing or otherwise adapting the design and functioning of their (…) recommender systems (…), so that they discourage and limit the dissemination of illegal content'.

Recital (62) specifies duties around disclosure of function on companies, requiring them to 'clearly present [to users] the main parameters for such recommender systems in an easily comprehensible manner to ensure that the recipients understand how information is prioritised for them'. It also requires companies to provide choices in algorithm design: companies should 'ensure that the recipients enjoy alternative options for the main parameters, including options that are not based on profiling of the recipient'. The requirement that users be allowed to choose recommender systems with no 'profiling' is particularly significant.

Recital (64) concerns access to company data about recommender system function. It says that regulators may 'require access to or reporting of (…) data on the accuracy, functioning and testing of (…) recommender systems'. This provision is probably sufficient to require companies to conduct the kind of fact-finding study that we advocate. The provision allows access to data on the 'testing' of recommender systems. But note that it doesn't allow regulators to ask for specific tests: it just concerns access to data reporting on tests that were conducted.

Within the Act itself:

Article (26) requires companies to conduct various 'risk assessments', in relation to risks of 'dissemination of illegal content', and 'intentional manipulation of (…) service'. These risk assessments should take into account how recommender systems may 'influence' these risks. This provision may be broad enough to allow regulators to ask for specific fact-finding studies to be conducted.

Article (27) requires companies to mitigate any risks that are identified. Mitigation measures include 'adapting' recommender systems. Note that these obligations are beyond the scope of our project: our project focusses just on identifying problems, not on what should be done to remedy them, as discussed in Section 3.5.1.

3.6.2 The UK's Online Safety Bill

The UK's Online Safety Bill was published in draft in May 2021, and has been subject to considerable parliamentary scrutiny since then. The bill includes a few provisions that appear to apply to recommender algorithms. Article 8 is particularly relevant. This clause sets out companies' obligations to conduct 'illegal content risk assessments': crucially, 8.5(d) includes a duty to assess 'the level of risk of functionalities of the service facilitating the (…) *dissemination* of illegal content [our emphasis], identifying and assessing those functionalities that present higher levels of risk'. The regulator (Ofcom) will have powers to impose 'service restriction orders'; Section 92 of the Bill notes that these include 'search engines which generate search results displaying or promoting content'. These provisions perhaps cover recommender systems.

The status of the bill is somewhat uncertain as we go to press: it was placed on hold twice in the last four months, apparently in response to the arrival and departure of successive new Prime Ministers. But the UK Government still appears committed to its passage.

3.6.3 US discussions about Section 230

In the wake of Frances Haugen's testimony about Facebook's algorithms to the US Congress, several bills were proposed in Congress that would require companies to be accountable for the decisions made by recommender algorithms (see a discussion here). These were all couched as amendments to Section 230 of the 1996 Communication Decency act, which famously exempts tech platforms from liability for the user-generated content they host. The amendments aim to exclude the actions taken by recommender algorithms from this exemp-

tion: Tom Malinowski's proposed Protecting Americans from Dangerous Algorithms Act is a clear case in point.

There is no immediate prospect of amendments such as these being passed, but it is interesting to see them being discussed. Note that the arguments reviewed in Section 3.5.1 about the difficulty of framing laws that hold companies accountable for 'harms' caused by recommender algorithms (see Keller, 2021; Llansó *et al.*, 2020; Thorburn *et al.*, 2022) very much apply to these amendments.

Section 230 has been the focus of court cases in the US, separately from discussions in Congress. Court of appeal judges in several states have been asked to rule whether the exemption from liability given by Section 230 for companies' hosting of content extends to their algorithmic recommendation of content. Court of appeal decisions have been gone both ways; a particular case, Gonzalez vs Google, was recently selected for hearing at the Supreme Court. The outcome is not yet known—but several commentators have suggested the issue is more appropriate for discussion by Congress than in the courts. Minimally, discussion by the Supreme Court will draw valuable attention to the question of what regulatory mechanisms are appropriate for social media recommender algorithms.

3.6.4 The US Platform Accountability and Transparency Act

The US Platform Accountability and Transparency Act (PATA) is a bipartisan bill before the US Congress, proposed by Senators Chris Coons, Rob Portman and Amy Klombuchar. The bill requires social media companies to allow independent vetted researchers to access certain types of company-internal data, to allow these researchers to 'release findings on the platforms' impact to the public'. The newly-formed Council for Responsible Social Media (Section 3.5.5) is advocating for this bill.

The PATA bill echoes certain provisions of the EU's Digital Services Act in the powers it grants vetted external agencies to access company-internal data. Note that in its emphasis on data access and transparency, it is not subject to the kind of difficulties that confront bils proposing amendments to Section 230 (as just discussed in Section 3.6.3).

Under PATA, independent researchers would be able to submit proposals to the National Science Foundation for studies that run internally to companies. PATA doesn't appear to give researchers the power to conduct new interventions on users, of the kind being trialled in the Toronto-Berkeley project (see Section 3.5.1). It just gives researchers access to company data (subject to certain privacy protections). Note that this power would be sufficient to allow the kind of fact-finding study that we are advocating for.

3.6.5 Proposals in the Netherlands State Commission

In the Netherlands, the Dutch State Commission on the Parliamentary System has proposed that an 'independent entity' should monitor platform recommendations, with a view to retaining 'diversity' and avoiding 'bias' (cited in Llansó *et al.*, 2020).

3.6.6 Voluntary transparency initiatives

We conclude by mentioning two nonbinding transparency schemes for tech platforms that have recently been announced. The first is an initiative from the OECD. The second is an initiative from companies and citizens' groups, developed to respond to communications from the EU Commission.

The OECD's Transparency Reporting Framework (VTRF)

The OECD VTRF (voluntary transparency reporting framework) is an international hub for submitting and accessing standardized transparency reports from online content-sharing services about their policies and actions on violent extremist and terrorist content (TVEC).

The reports are based on a questionnaire designed to be answerable by services of all sizes, which is intended to produce a baseline level of transparency. The VTRF was developed over a period of two years, with more than 100 of the world's leading authorities on addressing TVEC online, platform governance, and human rights—what the OECD calls 'the most extensive international multi-stakeholder consultation ever undertaken for a transparency reporting framework on TVEC'. Version 1.0 of the framework is supported by all 38 member countries of the OECD.

The VTRF is designed for several purposes: to improve the evidence base for informed policymaking, to increase the accountability of online platforms for increasing Internet safety while protecting human rights, to reinforce trust in the online environment, and also to reduce the costs for companies of transparency reporting, while increasing its efficiency and convenience.

The current VTRF questionnaire does not include questions about recommender systems, as far as we can tell. We suggest an updated version of the questionnaire should include specific questions about recommender systems.

The 'Strengthened Code of Practice on Disinformation'

Following a number of EU initiatives in the area of online disinformation (notably Commission communications on the EU Democracy Action Plan and on tackling online disinformation), many companies and citizens' groups have recently signed a 'Strengthened Code of Practice on Disinformation', which includes several specific commitments in the area of recommender systems.

In Section V(e), signatories 'acknowledge the significant impact that recommender systems have on the information diet of users, and therefore recognise that recommender systems should be transparent'. Measure 18.1 commits signatory companies to a specific action in relation to transparency: companies should 'publish the main parameters of their recommender systems'. (Commitment 19 suggests that the 'parameters' in question here concern the information sources that are used to 'prioritise or deprioritise' items in the content stream recommended by the system.)

In Measure 18.1, signatory companies commit to 'take measures to mitigate risks of their services fuelling the viral spread of harmful Disinformation'. These measures could include creating recommender systems 'designed to improve the prominence of authoritative information and reduce the prominence of Disinformation'. Companies also undertake to 'provide, through meaningful metrics capable of catering for the performance of their products (…including recommender systems …) an estimation of the effectiveness of such measures'. Metrics could include measures of 'the reduction of the prevalence, views, or impressions of Disinformation', and/or 'the increase in visibility of authoritative information'. Importantly, companies will 'insofar as possible, (…) highlight the causal effects of those measures'. A fact finding study with the form we advocate, but behavioural measures relating to disinformation, rather than TVEC, would probably fulfil these commitments.

The code of practice also includes commitments about allowing users various *choices* in the design of recommender systems. One choice is X. Another choice is detailed in Measure

22.2: companies will ‘give users the option of having signals relating to the trustworthiness of media sources [fed] into the recommender systems’. Alternatively, companies can commit to feeding these signals into their recommender systems without giving users a choice in the matter.

4 Thoughts on how to streamline discussions about recommender system transparency

In this final chapter, we take stock of the discussions about recommender systems we have been involved in over the past year. We reflect on which processes were effective, and propose a few ways in which discussions could be made more efficient. Our thoughts are primarily for the people who organise and participate in these discussions. But some of them may be helpful to GPAI experts discussing other areas of tech policy with governments and companies.

Sections 4.1 and 4.2 offer thoughts about how to have impactful discussions. Sections 4.3–4.6 offer thoughts about sensible groupings to create for discussions, and about dissemination of results from these discusisons.

4.1 Create high-level support for transparency initiatives

Important decisions in companies and governments happen at high levels, obviously. We have found that our transparency discussions were most effective when they could be understood as progressing commitments made at a high level by governments and/or companies.

It also seems to us (from our very limited experience) that high-level commitments of this kind are particularly effective when they are linked to future high-level meetings. As an example, the joint commitment made by the Prime Ministers of the UK and New Zealand to make an announcement about pilot projects on 'algorithms' at the Christchurch Call Summit (see Section 3.3) seemed to be very effective in generating discussion in government groups in the leadup to this Summit. It may also have played into the announcement of the Initiative on Algorithmic Outcomes at this year's Summit (see Section 3.4.2).

If high-level parties are committed to making an announcement about some project at a forthcoming meeting, then there is good impetus for doing the detailed work needed to define this project, and a clear deadline for completing the work. If there is no such commitment prior to the meeting, an agenda is still needed for the meeting. One thing that a technical group can do is to argue for certain topics to be placed on the agenda. This is another way of making progress at high levels. Of course, this requires knowledge of upcoming meetings, and awareness of the processes involved in organising them.

Summits that occur repeatedly are also a good mechanism for creating and progressing high-level commitments. Again, the Christchurch Call Summit is a good example: at one Summit, commitments could be made about work to be completed by the next Summit. Commitments are often more open-ended. For instance, there is no projected completion date for the privacy-enhancing technologies that feature in the Initiative on Algorithmic Outcomes. And in the Christchurch Call's Algorithms workstream, the processes committed to are only vaguely defined, and do not have a clear completion date. Commitments with fixed completion dates may be more helpful.

At a lower level, some of the multistakeholder discussions we have been involved in have led to 'recommendations'. For instance, the GIFCT's recommender system project made some

recommendations, that clearly required some level of company signoff. Even though these come with fixed completion dates, it's not clear that companies feel committed to following these recommendations.

4.2 Improve awareness about the processes of government amongst GPAI experts

The first year of our project on recommender systems was largely technical, and engaged us as AI researchers, in literature reviews and study designs. The second year, reported on in this document, has largely been an exercise in advocacy, with companies and with governments. In the first year, we developed a particular technical proposal about how best to study recommender system effects. In the year just completed, we have essentially been lobbying for this proposal.

Other GPAI projects in our working group have had a similar trajectory: for instance, the AI for public-domain drug discovery and Responsible AI for the Environment projects both developed technical proposals and then had to advocate for them.

GPAI experts are selected mainly for their technical expertise, but advocacy requires a very different skillset: it would be useful if GPAI's processes provided better methods for liaising with governments. The recommendations in our working group's Multistakeholder Expert Group Report around SDG accountability tools go some way towards addressing these needs. But we think more concrete assistance with the process of engaging with governments might also be helpful. In particular, we think the regular participation of delegates from relevant government groups in GPAI projects would provide some of the needed expertise in government mechanisms.

4.3 Involve company engineers in transparency discussions

Our remaining thoughts focus on the content of discussions on recommender system transparency, and on how they are organised.

A first suggestion here is that company engineers should play a far larger role in discussions. Company engineers and data analysts have vital expertise and insights to contribute to discussions about transparency methods with external researchers. They are the experts in the methods currently deployed for their platforms. But companies are very reluctant to make their engineers available for such discussions. It should be possible for external researchers to have discussions with company engineers about possible transparency mechanisms. External researchers may be needed to sign non-disclosure agreements, but most would be willing to do so.

Conversations with engineers are also necessary to develop privacy-enhancing technologies that are fit for purpose. We foresee the need for *ongoing discussions* about privacy-enhancing technologies. As companies' technologies change, and as new questions arise about user effects, the functionality of PET mechanisms is likely to need updating. PET technologies will have to keep pace with platform technologies, and with public discussions. Conversations with company engineers about the design of PET systems will be needed on an ongoing basis.

We also want to emphasise that privacy-enhancing technologies are in no way a substitute for discussions with company engineers. PET methods provide external researchers with

certain abilities to engage with companies' internal systems. But they don't provide any expertise or experience in the use of these methods. Companies have valuable expertise in this area, which can inform discussions about study design. Even with established PET methods in place, there should still be a forum for external researchers to engage, under suitable non-disclosure arrangements, with company engineers, to find good ways of answering questions in the public interest. These questions are primary: PETs are a valuable tool for helping answer these questions, but the key focus should be on the questions to be answered.

4.4 Focus discussions on concrete pilot studies

A second suggestion is that discussions on recommender system transparency should focus on fully specified pilot studies that could be conducted inside particular companies. The pilot studies proposed in the GIFCT's Technical Approaches Working Group (see Section 3.2.2) are a case in point.

A first step would be to specify some *possible* pilot studies. This process is one where company engineers could play a valuable role. They can assess proposals for feasibility—and also possibly for resourcing and/or cost. But they can also make proposals of their own about the best mechanisms that could be used to surface facts about the operation of recommender systems.

Once some technically feasible pilot studies have been proposed, a separate round of discussions can be held about the legal implications of these pilots. Some studies might raise more questions than others; sometimes technical modifications might be suggested that would allay certain concerns. But in all cases, legal discussion could be focussed on maximally concrete transparency exercises.

We are assuming here that these engineering and legal discussions happen inside *particular companies*. The issue arises as to how to compare the running of pilot studies across companies. This is certainly an important issue to discuss—and certain forms of study might be more transferrable between companies than others. It will certainly be important to define transparency measures in a way that applies to companies generally. But we think pilot studies running in particular companies will be useful in framing the more general definitions. We emphasise that the studies in question are *pilot* studies: their role is to further transparency discussions, rather than to surface results in their own right.

4.5 Better interactions between cooperative and regulatory discussions

A key question that arises in many groups but doesn't properly have a home, concerns how cooperative discussions about transparency mechanisms (such as those at GIFCT and the Christchurch Call) relate to discussions about regulation. We certainly feel that these cooperative interactions can usefully *inform* discussions about regulation. The whole point about the 'pilot studies' mooted at GIFCT (Section 3.2.2) is that they *try out* certain transparency mechanisms. If they work well, then some general statement of them can perhaps be built into regulations. Certainly, in the absence of trials of this kind, legislators have little option but to give regulators wide powers to access (and perhaps intervene in) company processes. The powers granted to regulators in the EU's Digital Services Act (see Section 3.6.1) are perhaps of this kind.

But we don't want to suggest that cooperative discussions are redundant if regulation is already under way. For instance, if regulations do grant wide powers of access to companies,

as in the case of the EU's DSA, the question of how regulators should *use* these powers is still a very important one.

Whether discussions about concrete transparency mechanisms happen prior to legislation, and inform it, or happen after legislation, and concern how it should be implemented, we foresee an *ongoing discussion* about the technical details of transparency mechanisms. Companies are always modifying their systems, and new questions about transparency mechanisms are likely to arise frequently.

We believe it would be useful to establish a *regulatory body* that serves as an intermediary between companies and governments, that is tasked with designing and piloting the necessary transparency mechanisms. The kind of ad-hoc discussions currently taking place about recommender system effects in the Christchurch Call and GIFCT, should in due course be coordinated by this regulatory body.

The exact role of a regulatory body in this area is of course a matter for a great deal of debate in itself. Should it be a collection of national regulatory agencies? (If so, how would they interact, to regulate social media platforms that operate across borders?) Or should it be an international body of some kind? (If so, what agency would legitimise it? How would it be governed?) We have no answers to these questions. Our main argument for a regulatory body is just that we foresee *a lot of work*, of an ongoing nature, in this general area: so some framework is needed to structure the work that will have to be done.

4.6 Create a public science around recommender system effects

A final thought relates to the above discussion of tech companies and governments. A tech company is *like* a government in some ways—in particular, in the wide-ranging effects of its decisions on large populations. 'Tweaks' to company algorithms are in some ways like 'tweaks' to government policies: for instance, changes in the minimum wage, or tax rates. Some of the relevant economic tweaks are conducted externally to government: for instance, central banks often set a country's interest rates. All these tweaks have large cumulative effects on the population. Similarly, the tweaks made by a tech company on its recommender algorithm can be expected to have large cumulative effects.

But there is an important difference here. The tweaks made by governments and central banks are informed by a whole discipline—macroeconomics—that studies the effects of economic policy changes. But there is no comparable body of public science that informs decisions about changes to recommender systems. At best, there are insights gained by engineers and data analysts within that particular company. For some companies, we guess there are minimal insights of this kind. What's needed is a *scientific discipline* devoted to the study of recommender system effects, that is a normal part of public science. This science would inform decisions made by all companies—and would also inform the operation of a regulatory body that determines and implements transparency mechanisms for companies in this area.

References

Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H., Patel, S., … Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 acm sigsac conference on computer and communications security* (pp. 1175–1191).

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).

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.

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. (2021). *Estimating the effect size of moral contagion in online networks: A pre-registered replication and meta-analysis.* OSF Preprints.

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.

Brown, M., Bisbee, J., Lai, A., Bonneau, R., Nagler, J., & Tucker, J. (2022). *Echo chambers, rabbit holes, and algorithmic bias: How YouTube recommends content to real users.* Available at SSRN 4114905.

Chen, A., Nyhan, B., Reifler, J., Robertson, R., & Wilson, C. (2022). *Subscriptions and external links help drive resentful users to alternative and extremist YouTube videos.* arXiv:2204.10921v1.

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, *1*(11), 769–771.

Dinir, I., & Nissim, K. (2004). Revealing information while preserving privacy. *Proc. ACM Symp. on Principles of Database Systems (PODS)*, *pp. 202-210*.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. *Theopry of Cryptography, Springer Lecture Notes on Computer Science*, *3876*(pp. 265-284).

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. foundations and trends. *Theoretical Computer Science*, *9*(3–4), 211–407.

Eckles, D., Karrer, B., & Ugander, J. (2016). Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, *5*, 20150021.

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, *80*(S1), 298–320.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … Vayena, E. (2018). AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689–707.

Geppert, T., Deml, S., Sturzenegger, D., & Ebert, N. (2022). Trusted execution environments: Applications and organizational challenges. *Frontiers in Computer Science*, *4*, 930741.

GPAI. (2021). *Responsible AI for social media governance: A proposed collaborative method for studying the effects of social media recommender systems on users.* Global Partnership on Artificial Intelligence report. (Authors: Knott, A., Hannah, K., Pedreschi, D., Chakraborti, T., Hattotuwa, S., Trotman, A., Baeza-Yates, R., Roy, R., Eyers, D., Morini, V. and Pansanella, V)

Huszár, F., Ktena, S., O'Briena, C., Belli, L., Schlaikjer, A., & Hardt, M. (2022). Algorithmic amplification of politics on twitter. *PNAS*, *119*(1), e2025334119.
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., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of SIGIR*.
Keller, D. (2021). *Amplification and its discontents: Why regulating the reach of online content is hard.* Knight First Amendment Institute Occasional Papers, Columbia University. Retrieved from https://knightcolumbia.org/content/amplification-and-its-discontents
Kemp, S. (2022). *Digital 2022: Global overview report.* DataReportal. Retrieved from https://datareportal.com/reports/digital-2022-global-overview-report
Kenig, B., & Tassa, T. (2012). A practical approximation algorithm for optimal k-anonymity. *Data Mining and Knowledge Discovery*, *25*, 134–168.
Konečnỳ, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., & Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.
Lada, A., Wang, M., & Yan, T. (2021). *How machine learning powers Facebook's news feed ranking algorithm.* Engineering at Meta report. Retrieved from https://engineering.fb.com/2021/01/26/ml-applications/news-feed-ranking/
Ledwich, M., & Zaitsev, A. (2019). Algorithmic extremism: Examining YouTube's rabbit hole of radicalization. *arXiv preprint arXiv:1912.11211*.
Llansó, E., van Hoboken, J., Leersen, P., & Harambam, J. (2020). *Artificial intelligence, content moderation, and freedom of expression.* Working paper of the Transatlantic Working Group on Content Moderation Online and Freedom of Expression. Retrieved from https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf
Mahato, G., & Chakraborty, S. (2021). A comparative review on homomorphic encryption for cloud security. *IETE Journal of Research*. Retrieved from https://www.tandfonline.com/doi/full/10.1080/03772063.2021.1965918
Morais, E., Koens, T., van Wijk, K., & Koren, A. (2019). A survey on zero knowledge range proofs and applications. *SN Applied Sciences*, *1*, 946.
Nick, E., Cole, D., Cho, S.-J., Smith, D., Carter, T., & Zelkowitz, R. (2018). The online social support scale: Measure development and validation. *Psychological Assessment*, *30*, 1127–1143.
Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, *3*, 96–146.
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).
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.
Rocher, L., Hendrickx, J., & de Montjoye, Y.-A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, *10*, 3069.
Rowa, J. (2022). *The contextuality of lone wolf algorithms: An examination of (non)violent extremism in the cyber-physical space.* GIFCT report.
Scharkow, M., Mangold, F., Stier, S., & Breuer, J. (2020). How social network sites and other online intermediaries increase exposure to news. *PNAS*, *117*(6), 2761–2763.
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.
Stray, J. (2020). Aligning AI optimization to community well-being. *International Journal of*

Community Well-Being, *3*(4), 443–463.
Stray, J., Halevy, A., Assar, P., Hadfield-Menell, D., Boutilier, C., Ashar, A., … Vasan, N. (2022). *Building human values into recommender systems: An interdisciplinary synthesis.* arXiv:2207.10192.
Thorburn, L., & Stray, J. a. (2022). *What will "amplification" mean in court?* Tech Policy Press: Technology and Democracy. Retrieved from https://techpolicy.press/what-will-amplification-mean-in-court/
Thorley, T., Llansó, E., & Meserole, C. (2022). *Methodologies to evaluate content sharing algorithms & processes.* GIFCT Technical Approaches Working Group report.
Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.
Watson, A. (2022). *Social media news worldwide—statistics & facts.* Statista report. Retrieved from https://www.statista.com/topics/9002/social-media-news-consumption-worldwide/
Whittaker, J. (2022). *Recommendation algorithms and extremist content: A review of empirical evidence.* GIFCT Transparency Working Group report.
Wolfowicz, M., Weisburd, D., & Hasisi, B. (2021). Examining the interactive effects of the filter bubble and the echo chamber on radicalization. *Journal of Experimental Criminology*. Retrieved from https://doi.org/10.1007/s11292-021-09471-0
Working Group on Content-Sharing Algorithms, Processes, and Positive Interventions. (2021). *Content-sharing algorithms, processes, and positive interventions working group part 1: Content-sharing algorithms & processes.* GIFCT report.
Zhao, C., Zhao, S., Zhao, M., Chen, Z., Gao, C.-Z., Li, H., & Tan, Y.-A. (2019). Secure multi-party computation: Theory, practice and applications. *Information Sciences*, *476*, 357–372.