Evaluating recommender systems in relation to the dissemination of illegal and harmful content in the UK

Pattrn Analytics & Intelligence

July 2023

OXFORD UNIVERSITY INNOVATION
UNIVERSITY OF OXFORD
PATTRN.AI

A report by
PATTRN.AI
Pattrn Analytics & Intelligence
264 Banbury Road, Oxford, Oxfordshire, England, OX2 7DY

Acknowledgement
Contributors:
Professor Philip Howard, Principal Investigator

Lisa-Maria Neudert, Co-Principal Investigator

Evaluating Recommender Systems Team

Dr John Gallacher, Research Co-Lead

Luke Thorburn, Technical Analyst

Dr Bence Kollanyi, Research Analyst

Richard Kuchta, Research Analyst

Understanding Recommender Systems Team

Alex Hogan, Research Co-Lead

Dr. Rachel Armitage, Research Analyst

Jingyi Grace Gu, Research Analyst

The authors sincerely thank the subject-matter experts who generously contributed their time and expertise to this research. We extend a special appreciation to Professor Jonathan Stray for his invaluable contribution as an internal peer-reviewer.

About this report

In October 2022, Ofcom commissioned Pattrn Analytics & Intelligence (Pattrn.AI) to examine possible methods for evaluating the impact of recommender systems. Ofcom wanted to understand the different ways in which online services could test whether their recommender system design choices were likely to increase or decrease the likelihood of their users encountering illegal and harmful content. Ofcom also sought to understand the merits of these different evaluation methods, including their efficacy, costs and any ethical concerns related to their use.

This report details our findings.

Table of contents

Table of figures

1 Glossary

AI Act – A new act proposed by the EU Commission, which would introduce new obligations for those building and using AI systems.

Adtech – Software and tools used to buy, manage and analyse digital advertising.

Amplification – The relative promotion of content on a service, which can be influenced by how a recommender system has been designed.

Borderline content – This is harmful content that is not judged by in-scope services to be illegal or otherwise violative of their T&Cs, but which is very close to the threshold. Services will often make a decision to down-rank this content or otherwise make it less visible to users.

Content Moderation – The process of detecting content that is irrelevant, illegal, harmful or otherwise considered undesirable, and removing or otherwise decreasing the visibility and reach of such content.

Digital Services Act or DSA – New legislation introduced by the European Commission that aims to create a safer online environment. The DSA came into force in November 2022 and in addition to creating new rules, updates the eCommerce Directive 2000.

Domain – The subject matter or interest which provides context for the use of a system, e.g. music, film, etc.

Engagement – Engagement is a set of user behaviours, generated in the normal course of interaction with the platform, which are thought to correlate with value to the user, the platform, or other stakeholders. Engagement can contribute to recommender decisions.

Extremism – This term does not have an agreed, single definition in the literature reviewed as part of this research. Broadly speaking, however, Pattrn defines extremism as the belief that the survival of the ingroup is inseparable from some kind of direct or offensive action against the outgroup, brought about through a process of radicalisation, where individuals make increased preparation and commitment to intergroup conflict (Berger, 2018 and McCauley & Moskalenko, 2008).

Extremist content – Content that expresses or promotes extremist views, as outlined in the definition of extremism above. This working definition is not within the scope of regulation under the Online Safety Bill.

Extreme content – This term does not have one definition agreed in the studies we reviewed. Where it is used in this report, it refers to content considered outside of and unacceptable to the mainstream because of the potential harm it might cause or incite. As such, extreme content as referenced here may in some cases include content considered illegal under the provisions of the Online Safety Bill, but this was not made explicit in the studies we reviewed.

Filter bubble – Describes the narrowing of content that is recommended to users, such that content feeds become homogenous and lack variety. Also often referred to as an echo chamber.

Freedom of Expression – In this report we are referring to the UN's Universal Declaration of Human Rights, Article 19, which states that everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers.

Generative AI – Algorithmic tools that can be used to create new content – including audio, code, images, text, simulations and videos – using simple prompts and instructions.

Harmful content – For the purposes of this paper, we use the term harmful content to refer to content that is in some way damaging to the users who view it. While some of this content may be illegal, some of it will not (e.g. online bullying or misinformation content). At the time of drafting, the Online Safety Bill does not explicitly define harmful content. Any reference to such content in this paper does not distinguish between content that is harmful to adults or children, and rests on common conceptions of what is harmful, which the authors have determined from their reading of online safety research.

Illegal content – References to illegal content in this paper refer broadly to content that may be illegal (but not determined as such to a criminal law standard or the standard required by the Online Safety Bill). Such references include, but are not limited to, content related to the priority offences listed in the Online Safety Bill, such as terrorism offences, child sexual exploitation and abuse (CSEA) and a number of others including fraud, hate crime and public order offences, encouraging or assisting suicide, the unlawful supply of drugs, and the purchase and sale of prohibited firearms.

Large Language Models (LLMS) – Language models are algorithmic systems able to generate text. Large Language models are trained on a large volume of text data – on the scale of tens of gigabytes – and generate text based on user prompts.

Online Safety Bill or OSB – A UK bill that will introduce new duties for user-to-user services and search services.

Rabbit hole – The process of recommending ever more extreme content to users over time, which may occur as a result of users engaging with that type of content in the past. Particularly likely among users who already exist in filter bubbles (see above).

Recommender system – An automated tool that interfaces with a library of content hosted on a digital platform to surface specific content for users. The system accomplishes this goal by analysing a number data points, which may include information about users and the properties of content.

Search engine – Software and tools designed to enable the organisation of digital information, such as URLs, which has typically undergone indexing and which surfaces indexed information in response to user input in the form of a search query.

TVEC – Terrorist and violent extremist content.

User-to-user (U2U) services: The OSB defines U2U services as an internet service where content that is generated directly on the service by a user or uploaded to or shared on the service by a user, may be encountered by another user, or users, of the service. Most commonly, when the term U2U service is used in this report, it is a reference to social media platforms.

Violative View Rate (***VVR***) **–** The estimated percentage of views on a platform that is of content that violates that platform's policies.

2 Executive summary

Recommender systems allow users to find content they enjoy and wish to engage with. Without them, it is hard to imagine being able to navigate the huge volume of information that is now available online. Yet the way they are designed can influence the degree of risk posed to users, with some design decisions resulting in users being more likely to encounter illegal and harmful content (see Box 1). This report looks at how online services might evaluate their own design decisions, giving them the insights they need to strengthen user safety in the process of building and maintaining their recommender systems.

Drawing on the information gathered through desk research, expert interviews and workshops with data scientists and engineers, we identify a three-part typology of evaluation methods:

- **Observation methods** – These involve collecting and analysing data about the type of content that users see. One example is the use of prevalence metrics, such as "violative view rate", whereby platforms count the incidence of harmful content being shown to users, as a proportion of all content that is recommended to them.

- **Experimentation methods** - These involve the manipulation of one or more variables in the design of recommender systems, to understand if/how those changes impact user exposure to illegal or harmful content. The best-known example of experimentation methods are A/B tests, which are performed in live environments. Some services also perform experiments using "web enabled simulation" tools that emulate the structure of those platforms, with automated bots that mimic the behaviour of real users.

- **Self-report methods** - These involve asking users to share information about the type of content they have recently seen on a service. One example is the use of experience sampling, where a number of users are prompted at regular or random intervals to describe the content they have recently seen on their recommender feeds. Another is the use of qualitative user diaries.

A key advantage of experimentation methods is that they allow for "**causal**" patterns to be observed, meaning that services can see how a particular design choice influences (or could influence) the extent to which users are exposed to illegal or harmful content. In contrast, the observation and self-reporting methods are largely "**descriptive**" in nature, meaning they cannot directly attribute observed outcomes to a particular variable.

It is still possible, however, to glean insights from these methods, by comparing those observed outcomes against any changes that are being made to the design of a recommender system (e.g. by surveying users before and after a major design change).

No evaluation method is perfect. While some may generate a wealth of insights, they can also entail significant costs for the services who deploy them. To perform A/B tests, for example, services would need to collect and store significant amounts of data, as well as to construct testing platforms to execute those tests and analyse the results. Services would also require skilled data scientists and data engineers to design and oversee the implementation of such tests.

While our research suggests that the largest services have the wherewithal to undertake these more demanding experimentation methods, this is unlikely to be the case for the smallest services, who may find that observation and self-reporting techniques are better aligned to their resource constraints. The smallest services may also be able to draw on external support to conduct these evaluations, making use of third party algorithmic assessment providers.

Alongside weighing up the costs and resource demands of different evaluation methods, services should also bear in mind their ethical implications. We find that the use of some techniques can pose risk to users, such as by compromising their privacy (e.g. through the collection and storage of personal data) or by deteriorating their user experience. Using the sock puppet method, for instance, could result in real users interacting with fake accounts, creating a disingenuous online experience.

Regardless of the methods deployed, services can improve the quality of an evaluation by following several principles of best practice:

- **Plan ahead** - Some evaluation methods require access to historic data relating to content recommendations and user interactions, potentially spanning several months if not years. Collecting this data well in advance is vital to the successful deployment of such methods.

- **Know the system architecture** - To be able to evaluate a recommender system, services need to understand the different components of that system and how it operates in practice. MLOps (machine learning operations) procedures can help with the organisation of evaluations.

- **Keep evaluating** - Evaluation is not a one-off event. As services change, and as the design of their recommender systems evolves, so too do the risks facing users. Services should therefore think of evaluation as an ongoing exercise and look to replicable methods to observe changes in outcomes for users over time.

- **Invite outside scrutiny** - Where appropriate, services should consider allowing independent experts to observe or directly undertake evaluations and audits of their recommender systems (e.g. as Oracle is doing for TIkTok), which would help to minimise potential bias in the design of those evaluations.

The field of research and practice in the evaluation of recommender systems is still nascent, and there are several gaps in our collective understanding of the merits of different approaches. However, we are seeing a number of promising developments that indicate a maturing ecosystem. This includes the introduction of new standards (e.g. ISO AWI 42005 which will provide guidance for AI system impact assessments), as well as collaboration efforts in civil society (e.g. Deb Raji's Algorithmic Audit Network, which is a forum for researchers and policymakers to share best practice). We look forward to seeing the outcome of these initiatives and the continuing development of this ecosystem in the months ahead.

Box 1: Investigating the impact of recommender systems

In addition to examining practical methods for evaluating recommender systems, Pattrn undertook research to investigate the impact of recommender systems on user exposure to harmful and illegal content, drawing on past investigations by academic researchers, civil society groups, and journalists. Our headline findings:

- Recommender systems can take infinitely different forms. They vary according to the information signals they process, the predictions they make about content, the weighting applied to those predictions, and the way ranked content is "re-ranked", among many other factors. No two systems are the same.
- This means there is little value in asking the question of whether recommender systems in the round cause harm to internet users; it is unhelpful to make generalised statements about a technology that can be deployed in a multitude of ways.
- It is more useful instead to talk in terms of *design choices*, and the extent to which specific choices increase or decrease risks to users. This framing also demonstrates that it is within the power of online services to make changes to their systems to reduce risk to users, without necessarily removing those systems in their entirety from sites or apps.
- The field of research examining the impact of these design choices is still nascent. It is also hampered by researchers having limited access to platform data. However, several academic, civil society and journalistic investigations have shown that the design of recommender systems can impact user exposure to harmful content (e.g. Whittaker et al., 2021; Water and Postings, 2018; and Ribeiro et al., 2019).
- Research has also shown that design choices can influence the extent to which users are led on "pathways" from benign to increasingly harmful content (also known as "rabbit holes"). Studies have shown that these effects can increase user exposure to a number of harmful content types, including self-harm content, eating disorder content and extreme content (e.g. Center for Countering Digital Hate, 2022; Whittaker et al., 2021)
- However, design choices are not the only factor shaping user exposure to illegal content. Some researchers, for example, have argued that rabbit hole effects occur more frequently in cases where users are already inclined to seek out harmful content.
- Design choices can also determine the likelihood of recommender systems being gamed by bad actors. This can happen if the design of a system is too simplistic (e.g. with only a small number of information signals feeding into its ranking decisions) or if granular details of the design are made publicly available.
- While it is possible to prove that design choices matter, it is harder to know in abstract *which* design choices matter the most. This is because a) there are a very large number of choices available to platforms and their engineers; and b) every platform has its own unique user base and content profile, which will determine the effects of a design change.
- Given that we cannot say with certainty that X design choice is riskier than Y design choice, it is up to online services themselves to evaluate the impact of their own design choices within their own contexts – a question that is explored further in this paper.
- Policymakers should continue to develop their understanding of recommender systems, learning more about the impact of different design choices and the methods services could use to better protect their users. In doing so, policymakers will need to be vigilant of more significant technological developments, in particular recent breakthroughs in generative AI, which could augment or one day supplant recommender systems as we know them.

3 Introduction

Recommender systems have become ubiquitous within user-to-user (U2U) services (e.g., social media platforms where users can share content), and it may be impossible to operate some services without them. These systems aim to curate and recommend the most relevant content for users and can provide a valuable business model to platforms. However, recommender systems have also been demonstrated to amplify harmful content to their users under certain conditions. Studies of social media platforms reveal that these recommender systems can increase the likelihood that a user will be exposed to a wide range of harmful material such as polarising viewpoints, misinformation, and even illegal material such as terrorist content (e.g. Fernandez et al., 2020; Milano et al., 2020; Whittaker et al., 2021). In addition, recommender systems can be vulnerable to malign actors who wish to abuse the system in order to manipulate user behaviour and promote illegal or harmful content (Zhang et al, 2019; Christakopoulou & Banerjee, 2019). Given that such abuse can have negative effects on users and on society more widely, it is important to effectively evaluate the impact of the specific choices that feed into the design of a recommender system. Such evaluation can increase our understanding of the comparative limitations of implementations of recommender systems.

To evaluate what assessment methods are available for U2U services and which could be used to increase online safety, this project provides a survey of methods and a consideration of the relative strengths and weakness of each. This project was commissioned by Ofcom, which will be appointed as the regulator for online safety in the United Kingdom under the Online Safety Bill and will take on new regulatory responsibilities once the OSB receives Royal Assent. As part of those functions, Ofcom may choose to recommend measures relating to the design of their recommender systems for the purpose of compliance with the safety duties of regulated services.

Recommender systems are used to curate content that users would consider relevant or useful, usually approximated in practice by content that generates engagement. Yet research has documented that such systems are frequently engineered to boost business metrics – such as click-through rates, user retention, ad revenue, or the time users spend on a platform – rather than optimising for different uses or values to the user. In order to increase such measures, recommender systems may promote content that evokes a strong response (Milano et al., 2020). This can inadvertently lead to users being exposed to illegal or harmful content, which can have negative consequences for both the users and the platform. We consider it is crucial for U2U services to strike a balance between optimising for these business metrics through the use of recommender systems, and ensuring user safety through the deployment of additional evaluation methods that take into consideration the broader context beyond user engagement. There are a wide range of approaches that can be used for this task, each consisting of a number of interacting methodological elements. These range from observational approaches which gather data and insights into how users interact with the system, to experimental approaches such as A/B testing which compare the impact of different versions of a recommender system on subsequent system behaviour.

This project aims to provide evidence-based insights on these assessment methodologies and practical guidance to Ofcom. This research used expert interviews and analysis of academic and grey literature to provide evidence-based insights on these assessment methods. In total, the team interviewed 31 experts from academia, civil society organisations, government, and industry (generating 20 hours of transcripts) and reviewed the wide variety of papers on the topic.

How do recommender systems work in user-to-user platforms?

A recommender system (or a recommender engine) is a type of information retrieval and ranking system that suggests content to a service user. Recommender systems are powered by a set of algorithms which, depending on what they are optimised for, select the content that is suggested to the user. The goal of a recommender system is to generate recommendations likely to be valuable to both the user and the platform, however, the exact metric or goal will vary. Examples of recommender systems considered for this research include those used within newsfeeds, video streaming, short-video sharing platforms, user-recommender, group/chat-room recommenders, and dating app recommendations.

Two historically popular approaches to building recommender systems are collaborative filtering and content-based approaches. Collaborative filtering systems recommend content to users based on what similar users have engaged with in the past (Ricci et al., 2011). The logic is that if person A and person B have a similar taste in a particular type of content, they may have the same taste in other types of content. In contrast, content-based systems recommend content to users based on the nature of the content that they themselves have engaged with in the past (Aggarwal, 2016). Unlike collaborative filtering systems, content-based systems analyse the features of an item of content itself such as the text or images it contains. The logic is that if a person has engaged with content exhibiting X and Y features in the past, they are likely in future to engage with other content that has the same features.

In practice, most platforms use some form of hybrid method that combines collaborative filtering and content-based approaches, along with a range of signals from other components of the U2U platform (Meserole, 2022). While the exact approach will vary across platforms, most large-scale recommender systems will follow the same basic steps: 1) take the inventory of available content, 2) filter out content that violates their content moderation policies, 3) select from this filtered list a set of candidate items the user is likely to be interested in, 4) rank the items in order of predicted interest for presentation to the user, and 5) partially re-rank items to create, for example, diversity among items that are ranked consecutively. A diagram of this typical workflow is given in Figure 1.

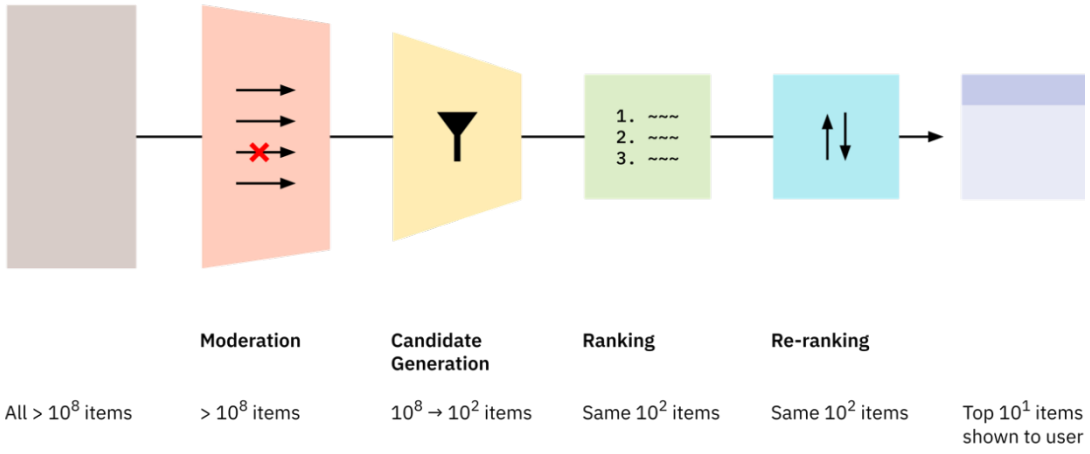

Figure 1 Workflow of the architecture of a modern recommender system, with item volumes typical of a large social media platform (reprinted from Thorburn et al. 2022).

Related work

The focus of this report is on the relationship between recommender system design choices and user exposure to illegal or harmful content, but most of the existing literature is not so precisely scoped. Academic work focuses heavily on fairness, privacy and the transparency (including interpretability and explainability) of recommender systems (Ge et al., 2022), while more recent works explore the impact on well-being, as well as legal and human rights (Stray et al., 2022). Researchers and practitioners in recommender systems actively investigate issues of the quality of a recommendation (Beutel et al., 2020); the exploration of quality already presumes that the content is not illegal nor harmful, and instead it is about whether the content will result in the user interacting with it, either through clicks or purchases (Zhou & Li, 2021). Only a few papers from this literature directly contend with how illegal content can be amplified by recommender models. That said, there is considerable work that looks at the relationship between a recommender system and a particular (variably defined) category of illegal and harmful content, from which useful and relevant insights can still be drawn.

Objectives and key research questions

The overall objective of this research is to develop insights into how actors in U2U services can assess the impact of their recommender systems and the role of those systems in the spread of illegal content. To achieve this objective, we structured our investigation under three primary research questions, each with related secondary questions.

Research Question 1: What are the key methodological strategies to evaluate the content decisions, systemic outcomes and impact on individual users of recommender systems across U2U services in the UK?

- Can these methods also tell us about which types of users are being exposed to illegal content?
- What methods could U2U services use to assess the extent to which recommender systems create pathways from harmful content to illegal content?
- Do different types of U2U service require different types of assessment?

Research Question 2: Comparing across different U2U services (by size, monetisation strategy, functionality, audience, etc.), what are the best practices for evaluating recommender systems currently used by industry?

- To what extent do online services already use these assessment methods?
- What do these assessment methods measure? How are these measures evaluated?
- How do these methods vary by efficacy and cost across different firms of varying size, monetisation strategy, functionality, and audience?

Research Question 3: How can U2U services adopt measures to evaluate recommender systems effectively and efficiently?

- Were online services to adopt assessment methods for the first time, what specific risks might these methods pose to user privacy?
- Is it feasible for smaller services to deploy these methods, as well as large ones? What in-house capabilities would be required to use them?
- How might online services act on the findings of any assessments they undertake?

4 Research design

The data collection and analysis for this project was completed over a three-month period from November 2022 to January 2023. Insights are drawn from three sources: (a) a literature review on assessment methods for recommender systems; (b) interviews with 23 experts drawn from academia, industry, government, and civil society organisations (CSOs); and (c) three discovery workshops with teams involved in the development and assessment of recommender systems, involving a total of eight experts.

The literature review was developed by searching for key terms in Google Scholar, SSRN, and arXiv. This combination of terms, amongst others, primarily included: "harms recommender systems", "content moderation recommender systems illegal content", "transparency recommender systems", "barriers responsible AI industry", and "evaluation recommender system". We consulted the proceedings from the ACM Conference on Recommender Systems (RecSys), Fairness Accountability and Transparency in Recommender Systems (FAccTRec), and the Workshops on Online Misinformation- and Harm-Aware Recommender Systems (OHARS). In addition to academic sources, we examined company transparency reports and reports from civil society organisations on methods to evaluate recommender systems from a responsible AI perspective. There are some limits to the approaches listed, however. Google Scholar, for example, depends on a ranking system that makes it harder to surface papers by scholars who are not already well established. Meanwhile, SSRN and arXiv, two of the most respected open access archives for social science and computer science research respectively, host unpublished work that has yet to go through peer review. Lastly, company transparency reports are subject to company influence and self-reporting biases.

Interview participants were identified using a combination of directed search by identifying those who have recently published work into the evaluation of recommender systems, and a snowball sample, by asking interviewees for the names of those who they believed would be useful in answering the research questions. In total across both approaches, we observed a 55% acceptance rate for interview invitations.

The distribution of interview and workshop participants across the affiliation categories was as follows: 38.7% associated with academia, 35.5% with civil society organisations, 54.8% with industry (or former industry), and 6.5% with government (note that participants can be affiliated with multiple categories, so the sum of these is greater than 100%). The Appendix contains a list of the interviews and workshops, including the pseudonyms that are used throughout this report, the affiliation type of the participants, and the date of the interview or workshop. When referencing a specific insight from an interview participant or workshop we use these pseudonyms.

Interviews followed a semi-structured approach to ensure consistency in themes and focus, while also allowing for a degree of flexibility to pursue specific lines of inquiry where relevant and permissible. The main lines of enquiry were: (a) to discuss the specific mechanisms by which recommender systems are implicated in the accessibility and distribution of illegal content, and (b) how user-to-user platforms could assess and measure this phenomenon. Questions then investigated best practice in this area across different services, and the barriers and limitations to successful assessment. Interviews took between 45 minutes and 1 hour to complete. A full list of research questions used within the semi-structured interviews is contained in the Appendix. In all cases, participants were informed of the goals of the research and gave their consent to participate in the interviews. Data was collected on a non-attributable basis, unless explicitly specified otherwise.

Following the data collection phase, the interview data was analysed using a general inductive approach where the main themes and topics were extracted over a series of intermediate summarisation steps, and the key insights extracted. Transcripts and interview recordings (where consent to record the interview was provided by participants) are contained within the Appendix.

In addition to the expert interviews, we also held three discovery workshops with teams involved with building and assessing recommender systems. One workshop was held with a team who worked at a social media platform, one with a team who worked at a news organisation, and one with a group of data science and machine learning engineers. These workshops also followed a semi-structured approach. In the case of the social media platform and news organisation, the discussion focused on the same questions as the expert interviews but as a more open and interactive discussion between all participants in the workshop. The data science and machine learning workshop followed a case study approach, where the participants were asked to imagine they were working for a social media platform and tasked with assessing the recommender systems in use (see the Appendix for details).

There are a few limitations to the study of recommender systems which should be noted. First, as recommender systems and the relevant data are proprietary, they cannot be readily accessed by the public or the research community for critical analysis and evaluation. Consequently, all external research into recommender systems exhibits blind spots, relies on assumptions or heuristics, and necessarily depicts a fragmented account of recommender systems in practical use. Second, and related to the first point, our sample of interviewees is balanced across experts from academia, civil society, and independent technologists, yet recruitment among current industry experts was less fruitful. While we spoke with a total of 17 experts with industry experience, the majority of these were either former employees of U2U services or experts working in adjacent industry areas rather than working directly for U2U services. As a result, insights from experts currently working within U2U platforms are underrepresented in this study, and we spoke directly to just two such platforms along with a media broadcast organisation. Third, the voices represented in this research – both of interview subjects and authors quoted – mainly represent European and North American perspectives, which is where many of the recommender systems that will become subject to regulation under the OSB have been developed. However, perspectives from experts as well as users in other regions are represented to a lesser extent in this research.

5 Analysis & findings

Recommender systems within U2U services are typically formed of complex “system of systems” and many interacting components, models, and decisions affect their design (CS1, CS2, CS3, DW1). This makes it challenging to assess their role in a real-world environment where user agency plays a significant role in their behaviour (Rogers, 2022).

As a result of this complexity there are a wide range of methodological elements available but currently no consensus on the single best approach or a clearly superior assessment strategy. These methodological elements range from more simple approaches which explore a user’s self-reported experience of a platform, up to complex simulations or experimental designs. The best methodology for a service to use will vary depending on the design of the service, the level of risk accepted, the resources available, and the type of harmful or illegal content to be assessed. It is likely that a combination of approaches may be appropriate for many services.

Challenges in assessing recommender systems

Additionally, when assessing the impact of a specific recommender system, it was repeatedly highlighted that it is important to choose a useful and realistic alternative against which to compare a specific design choice (DW1, DW2, CS3). Note, however, that it is not possible for recommender systems on social media to be truly value-neutral. Even a strictly chronological feed prioritises recency and directs attention towards frequent posters and those in nearby time zones (Lazar, 2023; Ovadya & Thorburn, 2023); as a result, causal assessment requires careful planning and design.

One point of tension regularly raised by experts is between the evaluation of recommender systems and the detection of illegal content – which is normally the remit of content moderation systems and teams (DW2, IND2). Many of the approaches to evaluating a recommender system require or assume that there is a “better” way to detect illegal or harmful content than what is actually used in the content moderation system (since otherwise the content found during the evaluation would not be on the platform in the first place). We discuss this tension later in Box 3, along with potential solutions proposed in 4.1 Fork B, but it is important to consider that the detection of this type of content is in itself a nuanced challenge.

Much of the academic literature on assessing the impact of recommender system design on social media focuses on methods that can be performed using publicly available data or through limited interactions with a service. These methods include using publicly available data from application programming interfaces (APIs), conducting user studies, or limited experimental studies using a small number of inauthentic “sock puppet" social media accounts which can be manually controlled or programmed to act in specific ways. However, platforms themselves will likely have access to a much wider array of data sources and assessment methodologies that are not publicly available. Researchers without access to platform data or the ability to conduct on-platform experiments often encounter difficulties when trying to understand how platform recommendations are personalised for different users. Using automated accounts to test a recommender system from the “outside” can be a slow and tedious process. This is due to the need to manually create the accounts while avoiding triggering spam or inauthentic account systems, and then design the behaviours which the accounts should follow. Conversely, U2U platforms have the ability to monitor the recommendations made on their platforms in real time (CS6). This disparity between the public debate and the internal approaches has the result of limiting the public debate and allowing for limited academic scrutiny of the currently adopted approaches for assessing the impact of recommender systems on social media. As a result of this gap, the academic literature on this topic tends to lag behind internal

platform research and platform changes (Whittaker, 2022), and the latest developments and findings of platforms may not be reflected in the academic literature for 2-5 years. This highlights the need for increased public debate on the best approaches for using internal data and internal access to assess the impact of recommender systems on social media. By increasing this type of public debate, it is likely that the maturity of these approaches will rapidly improve.

The following sections provide an inventory of the methodological elements identified in this research from both the literature review and the expert interviews, along with the relative strengths and weaknesses and the consideration points of each element. We then discuss what best practice would look like for a range of U2U services before considering the barriers to successful adoption of these best practices and the consequences of doing so.

5.1 Developing a methodological toolkit

In this section, we document the methodological elements available to platforms for evaluating the design of recommender systems in relation to illegal and harmful content. **Methodological elements represent decisions made in the course of specifying a complete evaluation method.** The inventory of methodological elements is given in Table 1.

At a high level, we can state whether each methodological element is compatible with three broad approaches to evaluation: observation, experimentation, and self-report. This framework aligns with that proposed by the Global Internet Forum to Counter Terrorism (Thorley et al., 2022).

Observation methodological elements are defined by their use of passively collected real-world data from U2U services. This data is then analysed in order to better understand how recommender systems have performed, the role they have played in determining what users consume on the platform, and their subsequent online behaviours. In most cases, observation provides descriptive (rather than causal) insight, meaning that it cannot attribute outcomes to the design choices of recommender systems.

The second approach, **experimentation**, is distinguished by the role the researchers play in manipulating one (or more) aspects of the environment in order to gather insights into how the system works or the impact of various components relative to a control condition. By observing differential outcomes across such groups, experimentation can provide causal information about the impact of particular design choices.

The final approach, **self-report**, involves directly asking stakeholders (users, publishers, developers) about their experiences with the systems of interest. Typically, these approaches gather more subjective information and are often used to assess thoughts, opinions, and feelings rather than the type of quantitative behavioural metrics gathered via observation. Self-reported data is often less accurate or complete than data collected using more objective or systematic approaches.

The alignment of methodological elements into these three approaches is given in Table 1, noting that elements can be consistent with multiple approaches. In section 4.1.1 we describe in more detail each of these elements, what they do, and how they work. For the most part, the discussion of the relative strengths and weaknesses of each element is deferred to Section 4.2.

Table 1 Inventory of assessment method elements for evaluating recommender systems.

Ref.	Fork Methodological Elements	Observation	Experimentation	Self-Report
Baseline Decisions				
A	**Who performs the evaluation?**			
A1	First party	✓	✓	✓
A2	Second party	✓	✓	✓
A3	Third party	✓	✓	✓
B	**What is the variable or outcome to be measured?**			
B1	Prevalence	✓	✓	✓
B2	Virality	✓	✓	
B3	Pathways	✓	✓	✓
B4	Real-world outcomes	✓	✓	✓
C	**What type of insight should the method produce?**			
C1	Speculative			
C2	Descriptive	✓		✓
C3	Causal	✓	✓	✓
Process Decisions				
D	**How to identify illegal or harmful content?**			
D1	Machine learning classifiers	✓	✓	
D2	User reports	✓	✓	✓
D3	User surveys	✓	✓	✓
D4	Civil society reporting			✓
E	**How to do causal inference?**			
E1	On platform experiment (e.g. A/B testing)		✓	
E2	Off platform experiment		✓	
E3	Causal inference on observational data	✓		
E4	Recommender system debugging		✓	
F	**How to survey users?**			
F1	Survey instruments	✓	✓	✓
F2	Experience sampling	✓	✓	✓
F3	Diary studies	✓	✓	✓
F4	Stimulated recall	✓	✓	✓
G	**What simulation to conduct?**			
G1	Whole platform simulations	✓	✓	
G2	Sock puppet accounts	✓	✓	
G3	Functional testing	✓	✓	

Definitional Decisions

H	**How to quantify virality?**		
H1	Number of shares or reposts	✓	✓
H2	Structural virality	✓	✓
H3	Probability of being shared	✓	✓
H4	Epidemiological model	✓	✓
I	**How to quantify pathways?**		
I1	Recommendation maps	✓	✓
I2	Distance	✓	✓
I3	Learning of unwanted preferences	✓	✓
I4	Increasingly extreme content	✓	✓
I5	Explore/exploit trade-off	✓	✓

5.1.1 Structure

It may seem natural to think that there is a finite set of methods for evaluating recommenders in relation to illegal and harmful content, that these methods are alternatives or substitutes for one another, and that a platform can "pick one off the shelf" and use it to evaluate their own recommenders. However, this view is misleading because there are many different questions one can ask about the relationship between recommender systems and illegal or harmful content, and methods that answer different questions are not true alternatives. In addition, the set of distinct methods available for answering any one of these questions may be (practically) infinite, as the number of ways to combine different methodological elements increases combinatorically.

Instead, we think it is more helpful to think about the space of methods as a "garden of forking paths". When deciding on a method, a platform or external evaluator must make many decisions about what things to measure and how to measure them. Each of these decisions constitutes a fork that differentiates the final method used from others that are possible. To fully specify or use a method requires many such decisions, so it is difficult to compare methods in their entirety. However, at each of these forks, there is usually a much smaller number of paths that can be taken, and these alternate paths can be more meaningfully compared. For this reason, we aim in this section not to exhaustively enumerate a set of alternative methods, but to characterise the most important or consequential decision forks that are often encountered when specifying a method, and the key methodological elements available at each fork.

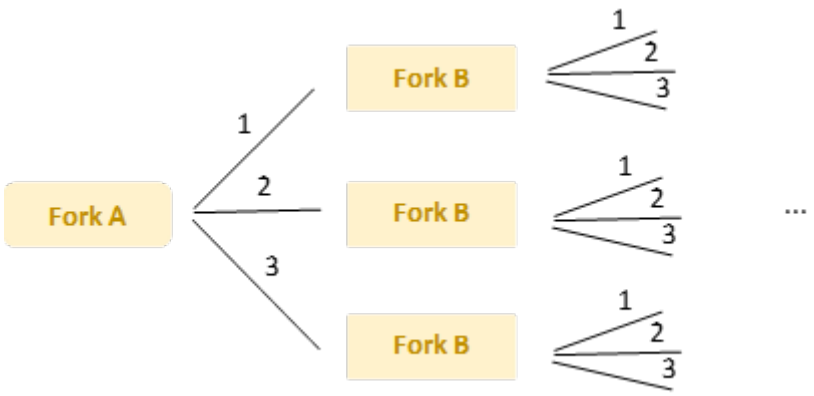

Figure 2 The beginning of the methodological "garden of forking paths".

The garden of forking paths is both a blessing and a curse. It is a blessing because (a) it provides flexibility to customise methods to a given context, trading off the strengths and weaknesses of different methodological decisions, and (b) it means that there are multiple, distinct methods that can be used in parallel, which is likely necessary to obtain a good understanding of the complex system of humans and machines in which recommenders operate. It is a curse because (a) it makes it more complicated to compare methods, and (b) it means that those who decide which methods to use have a lot of leeway, which can allow them to influence the ultimate findings of an evaluation. This last point has been raised in academia following concerns raised about *p*-hacking (the strategic selection of a data analysis method to produce statistically significant findings) and the replication crisis of many empirical scientific disciplines (Gelman & Loken, 2013). Note that the act of influencing the findings does not have to be deliberate or self-interested on behalf of the party performing the evaluation.

Below, we document nine common decision forks encountered when specifying an evaluation method, and for each fork, we describe the possible paths that might be taken, along with the most salient strengths or weaknesses, and any necessary prerequisites for choosing a particular option. Keep in mind that the separate forks combine into a much larger decision tree that maps the entire space of methods (Figure 2), but this tree is impractical to visualise in its entirety.

5.1.2 Decision types

Baseline Decisions

Fork A

Who performs the evaluation?

1. The platform (**first party**)
2. Someone that the platform contracts (**second party**)
3. Someone else (**third party**)

Arises in every method

First-party evaluation methods (Costanza-Chock et al., 2022; Raji et al., 2022) are performed by platforms themselves on their own recommender systems. First-party methods have the most access to platform data and methodological flexibility of any other type of method. As per instruction from Ofcom, first-party methods are the primary focus of this report.

Second-party methods (Costanza-Chock et al., 2022; Raji et al., 2022) are carried out on a platform's behalf by a person or organisation that they contract to do this. In theory, second-party methods have the same potential access to data and methodological flexibility as first-party methods, but in practice these advantages may be complicated by privacy or contractual challenges relating to the granting of system access to people who are not themselves employed at the platform. For example, Article 37 of the Digital Services Act (DSA) requires some platforms to contract second-party organisations to audit their compliance with other DSA requirements (Meßmer & Degeling, 2023). TikTok's hiring of Oracle to audit their recommender system (Fischer, 2022) is another example of a second-party audit.

Third-party methods (Costanza-Chock et al., 2022; Raji et al., 2022) are carried out by a person or organisation that has no official relationship or agreement with the platform, such as academics, civil society organisations, or (in some cases) regulators. Relative to other methods, they are the most limited in their access to data and must make do with what the platform makes publicly available – either via scraping (e.g. Hernandez-Suarez et al., 2018) or an API – or what users of the platform share with them – either as study participants (Knijnenburg & Willemsen, 2015) or by making data donations (Araujo et al., 2022; Boeschoten et al., 2020). It is also not possible for a third party to conduct on-platform experiments, which limits the extent to which third-party methods can infer causal relationships in a way that is ecologically valid.

While we focus on methods available to first parties in this report, note that this does include all methods available to second and third parties. Also note that independent reproducibility and verifiability is a desirable property for a method to have (Hutton & Henderson, 2015; Srivastava &

Mishra, 2023), which may justify the choice of a method that is available to third parties, even if it is the platform itself that carries out the work.

Fork B

What is the variable or outcome to be measured?

1. The **prevalence** of illegal or harmful content in a given context
2. The speed and/or **virality** of illegal or harmful content – see Fork H
3. The presence of **pathways** that lead users to illegal or harmful content – see Fork I
4. **Real-world outcomes** of illegal or harmful content on users or society

Arises in every method

Methods measuring **prevalence** aim simply to count the number or proportion of times that a particular type of illegal or harmful content appears in a given system or context on a platform. Examples include the prevalence of illegal or harmful content among:

- Content present at a URL on platform servers
- Content able to be seen by users if they search for it
- Content shown unprompted to at least *n* users, for some number *n*
- Content linked to but not hosted (e.g. at a URL posted in a comment)
- Content users remember having seen

Methods measuring **virality** aim to quantify the speed and volume of content with which illegal or harmful content spreads on a platform. See Fork F for a list of approaches to quantifying virality, such as by using epidemiological models.

Methods measuring **pathways** aim to quantify the extent to which a recommender system is leading users from ordinary content to illegal or harmful content. Our use of the word "pathways" here is deliberately vague, and there are a number of different ways it could be interpreted and quantified. For example, a pathway might be a situation where a user is shown illegal or harmful content that they would not have otherwise encountered, or where a user is shown increasingly extreme content. See Fork G for a list of approaches to quantifying pathways, such as by creating "recommendation maps" or looking for sequences of recommendations involving increasingly extreme content.

Methods measuring **real-world outcomes** aim to quantify undesirable phenomena such as poor mental health or offline illegal activity that may be exacerbated by exposure to illegal or harmful content. Such outcomes can be measured without causal inference (e.g., rates of depression among platform users) or with causal inference (e.g., the change in depression rates that can be attributed to the use of the platform and its recommender systems). Cases where a recommender system may be altering the preferences of users (Carroll et al., 2021; Evans & Kasirzadeh, 2022; Krueger et al., 2020; Thorburn, Stray, & Bengani, 2022a) are also included here as real-world outcomes. In academic literature, studies of the effect of exposure to content on real-world outcomes is known as media effects research (Valkenburg et al., 2016).

Which of these outcome measures is best depends on what it is that we care about. Arguably, reducing real-world outcomes, such as terrorist activity, child abuse, and poor mental health, is the ultimate goal, so we may ideally aim to measure the extent to which recommenders are contributing

to those (and similar) outcomes. However, such causal inference can be difficult to achieve (Thorburn, Stray, & Benghani, 2022), in part because events like terrorist attacks are rare. Conceding that, the next most relevant set of outcome measures may be prevalence measures that focus on first-person user experience, such as the proportion of content *seen by users* that is illegal or harmful. This focuses on content that is actually seen, which is necessary for it to cause real-world outcomes. In contrast, other prevalence measures, measures of virality and measures of pathways are of interest only for instrumental reasons; they may indicate that illegal or harmful content is being seen more than it would under some counterfactual recommender system, and hence is more likely to cause real-world outcomes.

A discussion of which of these outcome measures are typically used by U2U platforms is provided in Box 2.

Box 2

Outcome measures used by industry

A typical large platform will be monitoring dozens or hundreds of metrics over time. If the trialling of a change or the addition of a new feature to the recommender system during an A/B test causes any of these metrics to cross pre-defined thresholds, a human review of the change may be required to ensure that changes are only deployed if they are seen as causing an overall improvement in the metrics being monitored.

Only some of these metrics are publicly available, and only some relate to recommender systems and illegal or harmful content. One class of metrics about which there is public information is that related to "violative" content. Violative content includes illegal and, in some cases, harmful content such as hate speech or bullying that violates a platform's content moderation policy. A review of metrics related to recommender systems and violative content that are known to be used by platforms is given in Table 2. Platforms may also test for other metrics that are not publicly known.

Currently, "companies have sole discretion to decide which metrics they report on, how they calculate the data they share with the public, and which metrics they do not report on" (Singh & Doty, 2021; T1; CS4). Some platforms, including Meta and Google, have published metrics related to "harmful" content as part of a self-regulatory initiative to combat disinformation online (Google, 2023; Meta, 2020; Twitter, 2021; TikTok, 2022)). While these efforts are important, the descriptions of these measures are broad and not standardised across platforms.

Table 2 Metrics known to be used by companies to measure the impact of their recommender systems in circulating violative (including illegal) content.

Platform	Metric
Meta	Prevalence, i.e., the percentage of all content views on Facebook (or Instagram) that were of violating content in a particular content category
Twitter	Number of removed Tweets that received fewer than 100 impressions.
	Number of removed Tweets that received between 100 and 1,000 impressions.
	Number of removed Tweets that received more than 1000 impressions.
	Percentage of total Tweet impressions that were on violative Tweets

TikTok	Percentage of videos removed for violating terms of service or community guidelines removed before receiving any views. Videos not eligible for recommendation in users' For You feed (Exclusive to Election related content).
YouTube	Violative View Rate, i.e., what percentage of views on YouTube comes from content that violates their policies (O'Connor, 2021).

Each of these metrics is limited in the insight it can provide. For example, consider the Violative View Rate (VVR) which is used by YouTube. The VVR is calculated by taking a random sample of videos, tasking human reviewers to label them as violative or non-violative, and estimating the percentage of views these videos collectively received that accrued to violative videos.

When A/B testing alternative recommender systems, VVR can be used to evaluate the impact of design choices on the prevalence of violative content among views of videos. However, VVR on its own conveys no information about the severity of the violative content. For example, child sexual abuse material may have a low VVR because it is only circulated among a small number of users, yet the fact that even one piece of this type of material appears on the platform should be cause for concern.

The sampling method used also matters; a simple random sample and a complex stratified sample will give different results. Beyond the YouTube example, illegal content may be more likely to be found in certain parts of a platform, such as groups, messages, or events. Conditioning on these contexts when sampling may be important to accurately estimate the overall prevalence of illegal or harmful content. Finally, because content is labelled as violative by human reviewers, their interpretation of the platform's policies will impact whether the content is considered violating or not. Research has found that human reviewers who label content frequently diverge in their interpretations of policies and can be influenced by their demographic characteristics, as well as those of the user who posted the content (Waseem, 2016; Waseem & Hovy, 2016). Platforms can mitigate these biases by including labellers from a wide array of demographics, reporting inter-annotator agreement metrics, and hiring experienced and well-compensated labellers for this task.

Fork C

What type of insight should the method produce?

1. **Speculative** or **predictive** insight about what might happen – see Fork G
2. **Descriptive** insight about what has happened
3. **Causal** insight about what can be attributed to a recommender – see Fork E

Arises in every method

Speculative methods, such as simulations or predictions, provide insight about what *might be*, not what *is* (e.g., "if assumptions *z* are true, then *x* people will see content *y*"). See Fork I for a list of simulations that investigate the relationship between a recommender system and illegal or harmful content, including the use of sock puppet accounts or functional testing.

Descriptive or observational methods provide insight about outcomes (e.g., "*x* people saw content *y*"), but not what caused those outcomes. These may involve simply "counting", or they may also involve running further descriptive statistical analysis on top of the observed data. For example, analysts may fit epidemiological models to estimate the virality of certain types of content (see Fork F) or use statistical models to understand associations between the demographics of a user and their exposure to illegal or harmful content.

Causal methods provide insight about the extent to which outcomes can be attributed to a recommender system (e.g., "the algorithm caused *x* more people to see content *y* than otherwise would have under counterfactual *z*"). They arguably provide the most insight about the degree to which a recommender is implicated in the spread of illegal or harmful content, relative to a clearly specified alternative scenario. However, causal methods can be difficult to perform well in highly interconnected online networks, and the findings may not generalise broadly (Thorburn, Stray, & Benghani, 2022). See Fork E for a list of approaches to causal inference in the context of recommender systems, including on-platform experiments, such as A/B testing.

All the outcome variables listed in Fork C can be studied in speculative, descriptive, or causal ways. Note that if only seeking descriptive insight of a prevalence outcome variable, there are arguably no more significant methodological forks. A platform must simply write some code to count the prevalence of the particular content in the specified context, using the chosen method for identifying false negatives. There may, however, be complexity in formalising what it means for content to have been "seen" and to aggregate those figures across multiple interfaces and recommender systems used by a platform.

Process Decisions

Fork D

How to identify illegal or harmful content?

1. Use **machine learning classifiers**
2. Rely on **user & other third party reports**
3. Rely on **user surveys** – see Fork F
4. Rely on **civil society reporting**

Arises in every method

Recommender systems on online platforms are usually integrated with some form of content moderation system. If there is illegal or harmful content being recommended, it is content that has not already been filtered out by existing content moderation classifiers – that is, content which represents false negative classification errors. A core challenge when evaluating the involvement of a recommender with illegal or harmful content is that one must be able to identify these false negatives, which existing classification systems have already failed to identify. Broadly, there are four methods available for identifying false negatives.

Machine learning classifiers (Gorwa et al., 2020) are statistics or machine learning algorithms that classify items of content as belonging to a particular class that has been deemed illegal or otherwise harmful by a platform. In order to identify illegal or harmful content that has not already been excluded by a platform's content moderation system – false negatives – such classifiers must be different from, and preferably more accurate than, those that are used for content moderation. These high performing models may be too computationally expensive to run continuously in production for all content uploaded to the platform, may have precision/recall trade-offs which make them unsuitable for use in content moderation due to their risk of false-positives, or may simply be so new that they are not yet implemented at production scales. There are also open-source and commercial classifiers available for some categories of content (e.g. Davis, 2019; Google, 2018; Jigsaw, 2021; Thorn, 2020), but these are not a complete solution; classifiers will likely be more accurate if they are trained on a platform's own dataset, and the cost of commercial classifiers can become prohibitive, especially if used at platform-wide scale. Datasets used to train these models consist of items of content paired with labels (e.g., "terrorist content", "child sexual abuse material", etc.) generated by human raters, who the algorithmic classifier learns to imitate. Typically, raters use some form of classification rubric (e.g. Davey et al., 2021; Holbrook, 2015, 2017). There is often disagreement among human raters, so it is best practice for labels to be chosen according to the consensus or majority opinion (e.g. Davidson et al., 2017; Gordon et al., 2022), though this practice itself overstates the accuracy of classification (Gordon et al., 2021). This labelling process requires significant effort, although many platforms will be accumulating some of this data as a by-product of their content moderation efforts.

User reporting functions are forms that allow users to flag content they have encountered as potentially illegal or harmful. This content can then be reviewed by a human moderator or used as a useful signal for future investigation (see Fork E4). The number of items reported by users can also be a useful metric used to evaluate changes made to a platform or service; an increase in the reporting of illegal and harmful content may indicate that a change has surfaced more problematic content (Ind5). It should be noted that this is a lagging indicator of harm however, as the content will

have already been seen by potentially a large number of users (Ind5), and also that users need to be sufficiently engaged with the platform to report content reliably and consistently. User reporting is also readily gamed by users who, for example, flag content authored by another user they disagree with or wish to harm reputationally. For this reason, platforms seldom take individual user reports directly at face value, but usually review user reports before taking action (Dwoskin, 2018).

User surveys are web forms used by platforms to elicit information from their users. These can take the form of questionnaires used to proactively ask users about their experiences on the platform and can be used to ask a user whether and to what extent they have been exposed to particular kinds of illegal or harmful content. Both surveys and reports are able to catch false negatives because they relate to content that is currently visible on the platform. See Fork F for a list of available surveying methods, including experience sampling and stimulated recall.

Civil society reporting by third-party academics, journalists, or civil society organisations can identify when recommendation of illegal or harmful content is occurring on a platform. Groups such as Tech Against Terrorism[1], the Institute for Strategic Dialogue[2], The Markup[3], and others are able to investigate both recommender systems and the groups that use them, and in doing so identify illegal or harmful content that has not been excluded by existing content moderation systems. In cases where malign actors are acting strategically to avoid detection, such as by using "algospeak" (that is, using euphemisms or code-words to circumvent existing classifiers (Lorenz, 2022)), false negatives may be most likely to occur along the frontier of new and newly strategic content. Civil society groups may have expertise about the strategies used by malign actors that does not exist within platforms themselves, and so help to identify false negatives.

Similar third-party reporting of illegal content to U2U services can also come from government and law enforcement agencies, such as Counter Terrorism Policing's Internet Referral Unit[4] or Europol's Internet Referral Unit[5]. These organisations use a combination of dedicated staff and crowdsourcing to detect and investigate malicious content online.

Given that U2U services will remove illegal content from their systems once it is discovered, these measures will be used to retrospectively evaluate the recommender systems in use on a platform and the specific design decisions that went into each, rather than form the basis of an ongoing assessment of how this particular content is recommended as it is left on the platform.

Box 3

The interplay between recommender systems and content moderation

The interplay between content moderation systems and recommender systems sits at the crux of how illegal content moves through and is amplified on platforms. Indeed, the Australian Government's Office of the eSafety Commissioner acknowledges that recommender systems are a reflection of the content moderation decisions within the platform: "as well as potentially amplifying harmful content or inappropriately targeting users, services' recommender systems are routinely exercising content moderation decisions, whether in order to throttle the reach

[1] See https://www.techagainstterrorism.org/
[2] See https://www.isdglobal.org/
[3] See https://themarkup.org
[4] See https://www.counterterrorism.police.uk/together-were-tackling-online-terrorism/
[5] See https://www.europol.europa.eu/about-europol/european-counter-terrorism-centre-ectc/eu-internet-referal-unit-eu-iru

of 'borderline content' for all users, or to direct particular kinds of content away from users who might be offended by it" (eSafety Commissioner, 2023). However, there exist few case studies that showcase this dynamic, largely because platform failures remain behind closed doors except for a few highly publicised incidents.

In 2021, Facebook (which rebranded to Meta in October 2021) came under scrutiny for a recommendation feature that suggested users "keep seeing videos about primates" in reference to a video of police brutality against a Black man that appeared on Facebook. Facebook took action to address the issue, as outlined in a statement from Facebook VP Tom Alison, shared publicly on Twitter. First, the topic recommendation feature was disabled, as well as "other features related to topical recommendations". Second, the company rallied resources to execute a root cause analysis to discover how the "primates" label was applied to the video in the first place.

Fork E

How to do causal inference?

1. Conduct an **on-platform experiment**
2. Conduct an **off-platform experiment**
3. Do **causal inference on observational data**
4. Perform **recommender system debugging**

Arises only in methods seeking causal information (Fork C)

Experiments are the gold standard for causal inference and involve the manipulation of some initial conditions and comparison of outcomes across treatment and control groups. A common form of experiment in the context of online platforms are **on-platform experiments** or A/B tests, which compare outcomes when users are exposed to different versions of the platform (e.g., different versions of the recommender system). The outcome measure of interest is recorded for each group and any observed differences can be compared using statistical hypothesis testing to check for a significant difference (Young, 2014).

Within the category of A/B tests there are a wide range of possible specific implementations and methodological decisions (James et al. 2013). This range of choices make a standardised approach to on-platform experiments a challenge. The simplest example is a split tests where a single change is compared across a treatment and control group, while more complex multivariable tests can also be performed where multiple changes (>2) are compared over an equal number of treatment groups. In multivariable tests, the treatment groups can be nested within one-another in order to help control for confounding variables, and in these cases more complex statistical tests are used in order to draw meaningful insights. At the most complex level A/B tests can be run dynamically over a wide range of treatment options where 'poor' performing conditions (according to the outcome measure of interest) are dropped in favour of better performing conditions. One way of doing this is through

a multi-armed bandit experiment which uses machine learning to dynamically increase the users allocated to better-performing conditions.

The advantage of this type of experiment is that poor performing conditions are dropped quickly, so there is no wasted time or monetary expense in continuing these branches. This means that a higher number of treatments can be tested compared to split testing (Slivkins, 2019). The limitation is that performing statistical tests on the outcomes of a multi-armed bandit experiment is not possible due to the dynamic nature of the treatment allocation, and so it is not possible to demonstrate an observed difference in two conditions is statistically reliable (Nishimoto, 2021). Notably, on-platform experiments cannot be performed by third parties, but they could be performed by second parties working in collaboration with a U2U platform if the platform is willing to provide access to platform design and data collection resources.

Off-platform experiments are experiments performed using non-platform infrastructure. This includes browser extensions (e.g. Kohlbrenner et al., 2022) which are installed by a participant and then measure (and in some cases, manipulate) what content they see online as well as how they interact with it. For example, a browser extension might instruct the participant to follow (or not follow) certain accounts, and then observe how that influences the content recommended to them. Other modes of delivery for off-platform experiments include online surveys, which are common in media effects research, or more elaborate mock-ups of platform interfaces. In all cases it is possible to perform a range of experimental designs. Unlike on-platform experiments, off-platform experiments can be performed independently by third parties, but because they are limited in the extent to which they can use native platform interfaces, the insight they generate is usually less reliable than that produced by an analogous on-platform experiment.

In some cases, it is possible to do **causal inference on observational data**, meaning that it may not be necessary to conduct an experiment in order to attribute changes in outcomes to a recommender system. There is a well-established set of methods developed within the computational social science literature for doing this, including using instrumental variables, differences-in-differences, regression discontinuity, subclassification, matching, propensity scoring and synthetic control methods (Cunningham, 2021). Versions of these have been developed within the recommender systems literature under names such as “off-policy learning” and “counterfactual evaluation” (Saito & Joachims, 2022).

Each of these has certain requirements on the structure of the underlying dataset to be validly applied. For example, if a platform has a time series of event data (such as click rates) and wants to measure the impact of a specific intervention at a given time but has no comparable control condition, it is possible to use causal impact analysis to generate a synthetic counterfactual control (i.e., how the response metric would have evolved after the intervention if the intervention had never occurred). This is done by modelling the difference between the original time series and other comparable timeseries data that is unaffected by the intervention, and then projecting this difference forwards (Brodersen & Hauser, 2017). This has been used to measure the impact of hostile interference in online conversations in the absence of formal control conditions (Gallacher & Heerdink, 2019).

In practice it will often be simpler and more defensible for a platform to simply conduct an A/B test than to attempt causal inference using observational data, though one important exception to this is cases where it would be unethical to perform the experiment (e.g., if experimentation would mean

deliberately exposing users to content thought likely to influence them into posting incitements to violence). In such cases, studies based on observational data may be the only ethical option.

Recommender system debugging (CS3) – also known as root cause analysis – takes inspiration from debugging in software development. In this method, one starts with known cases of illegal or harmful content being shown to users and works backwards through the recommender system, examining the inputs and outputs of different stages in the algorithmic pipeline to understand why the content was shown. This might involve removing or altering certain elements of the recommender to conduct mini experiments and test whether the content would still be shown under an alternative system design (known as ablation or degradation studies) (CS3). This analysis can then be generalised to a broader class of content that contains the item which originally motivated the debugging process. Detailed platform logging can aid investigations, including recording the content presented to users, as well as the intermediate scores which were generated for this content within the internal recommender system processing (e.g., the predicted probabilities of engagement or content moderation classifications which were used to promote or downrank the content).

Fork F

How to survey users?

1. With questionnaire-based **survey instruments**
2. Using **experience sampling**
3. Using **diary studies**
4. Using **stimulated recall**

Arises only in methods that use user surveys to identify illegal or harmful content (Fork D)

Survey instruments are questionnaires used to elicit particular information from users. Such survey instruments can be ad hoc (developed in-house at a platform) or standardised (taken from existing literature). Ad hoc survey instruments can be tailored to a particular situation or context, but may not be psychometrically validated (that is, shown to be reliable measures of what they are intended to measure) or comparable with existing literature or across platforms. In contrast, standardised survey instruments are better able to produce data that can be compared across contexts, and may have been shown to be a valid measure of particular outcome variables (Phellas et al, 2011). Among other examples, survey instruments have been used by academics to measure indigenous peoples' experiences of harmful content (Kennedy, 2020), by academics to rate the severity of different types of harmful content (J. A. Jiang et al., 2021), and by both academics and platforms to measure the relationship between social media use and social comparison (Burke et al., 2020; Jiang & Ngien, 2020).

The remaining three methods are specific surveying strategies intended to improve the reliability of data that is self-reported by users.

Experience sampling (Csikszentmihalyi & Larson, 2014; Hektner et al., 2007; van Berkel et al., 2017**)** involves prompting the user at regular or random intervals to complete a short survey – perhaps only a single question – about their current or recent experience on the platform. Of the three strategies here, it is probably the least burdensome for the user and the cheapest to implement, as the same survey can be asked each time in an automated way through a platform interface.

Diary studies (Hyers, 2018) involve asking users to keep a qualitative diary and log specific information over a period of time, which can then be analysed by researchers. If completed properly, diary studies can be burdensome for participants due to the time involved in keeping detailed records.

Stimulated recall (Griffioen et al., 2020) is a surveying or interview method that involves prompting users to elaborate on some form of objective record of their interactions with a platform. For example, the user may be shown the history of items recommended to them and asked to explain how they responded to each one, or why they engaged with some rather than others. Stimulated recall is likely more costly for a platform to implement than experience sample, as the questions asked must be personalised to each participant in the study.

Fork G

What simulation to conduct?

1. Perform **whole platform simulations**
2. Create **sock puppet accounts** with prescribed behaviour, and observe what they see
3. Perform **functional testing**

This list is not exhaustive.

Arises only in methods seeking speculative (rather than descriptive or causal) insight (Fork C)

Whole platform simulations of a social network can, if well calibrated to reality, be used to find bugs or vulnerabilities and preliminarily test new features without needing to resort to experimentation involving human users (Gausen et al., 2022). A major example of this is Facebook's WW simulation (Ahlgren et al., 2020; Harman, 2020; Vincent, 2020), which first trains thousands of models or more to emulate the behaviour of real users, and then lets these bots interact with each other using production platform code, albeit with no ability to communicate with or influence the experience of human users. Such simulations are labour intensive and for the most part could only be performed by large platforms, and the trained bots may not accurately simulate the behaviour of real users, particularly over the long term (see below). If ethical concerns are minor, it is more informative to use a simple A/B test on a new feature to see the effect it has on illegal or harmful content, rather than to simulate its implementation (DW2).

Sock puppet accounts are synthetic accounts created within a U2U service which can be programmed, or manually controlled, to follow certain patterns of engagement when interacting with a recommender system. The items of content that the system recommends in response can be logged and subsequently analysed (e.g. Haroon et al., 2022; Hobbs et al., 2021). This method can be used to estimate how common illegal or harmful content is from the user's perspective, assuming that users behave in a certain way. This approach has been used to observe how recommender systems can influence the behaviour of children within U2U services (5Rights Foundation, 2021). While this approach can generate unique and useful insights, it is most useful for assessing what happens when users first join a platform (the initial 10–20 or so interactions) and for measuring the first few recommendations (DW2). However, it is very difficult to create realistic experiments and fake account histories that reflect longer interaction with a platform, and the longer these types of studies try to simulate activity, the lower the validity of the findings they generate. As such, these

types of studies are criticised for failing to incorporate user agency into their assessment of recommender systems, and do not account for how users actually interact with the recommender systems in a real-world setting (Ribeiro et al., 2023). Including user-agency in these synthetic assessments is challenging, but it is proposed to help alleviate an apparent paradox in the academic literature where studies often draw opposing conclusions about the role of recommender systems and the promotion of illegal content (Brown et al., 2022; Hosseinmardi et al., 2021).

Functional testing or stress-testing (DW3) is a method which takes inspiration from the use of unit tests in software development. It consists of having a number of well-defined scenarios where vetted, gold-standard recommendations are known. These recommendations could consist of content that is deemed to be benign and non-violative and suitable for recommendation, or alternatively, these could be content which is known to be unsuitable for recommendation. A given recommender can then be tested by comparing its own recommendations to those defined in the test. This approach has been highly successful within the sphere of validating hate speech detection models (Röttger et al., 2021). Functional testing is considered a form of simulation because the scenarios represented in the tests are not necessarily reflective of the true distribution of scenarios encountered in practice.

Any method of simulation faces fundamental limitations with respect to ecological validity. There is no guarantee that formal models of user behaviour or the platform environment accurately model reality, so simulations can only show us what might be, not what is.

Definitional decisions

Fork H

How to quantify virality?

1. Use the **number of shares or reposts**, optionally per unit of time
2. Use **structural virality**, which quantifies "grass roots" vs. "broadcast" sharing patterns
3. Model the **probability of being shared**
4. Use an **epidemiological model**

This list is not exhaustive.

Arises only in methods seeking to measure virality as an outcome variable (Fork B)

Virality is a term used to describe the degree to which online content spreads easily and/or quickly across many online users. It is possible that recommender system design choices can influence the virality of illegal or harmful content, and so influence the number of people exposed to it. To evaluate this possibility, it is necessary to formally quantify virality using data about how such content has been shared or reposted.

Most simply, virality can be formalised as the **number of shares or reposts**, optionally per unit time (e.g. Brady et al., 2020; Bruni et al., 2012). Such a measure is impacted not just by the "virality" of the content but by the number of users on the platform and the popularity of the users who initially shared the content, so in isolation is not a good measure of virality. Another approach to quantifying virality is to use **structural virality**: the average distance between all pairs of nodes in the reshare graph (a tree structure) for a particular item of content (Goel et al., 2016). Intuitively, structural

virality attempts to quantify the degree to which content is propagated via a "grassroots" rather than "broadcast" pattern of sharing. Grassroots sharing is where community/individual sharing of content is the driver of dissemination, while broadcast sharing is where the content is disseminated to a large audience because they might have a large number of followers.

The above two approaches require minimal modelling, meaning that they can be simply "counted" or "read off" from a platform dataset. Without more extensive modelling however, such measures can be vulnerable to confounding. For example, posts from accounts with a large number of followers or subscribers may appear more viral simply because they are initially distributed to a larger audience, giving them a "head start" in propagating through the network. Modelling virality as a function of other predictor variables (e.g., the size of the first account to post the content) is important to control for such confounders. There are also more latent ways of quantifying virality which inherently require a model to estimate.

One such approach is to model the **probability of a given item of content being shared** by each user who is exposed to it (e.g. Berger & Milkman, 2012; Hansen et al., 2021), using simple statistical models such as logistic regression to model this probability as a function of whether an item of content belongs to certain categories of illegal or harmful content.

Similarly, **epidemiological models** can be used to model the spread of information through a social network analogously to the spread of disease through a physical population. In this approach, virality is generally formalised using an estimate of the expected number of users each exposed user will go on to share the content with (whether voluntarily or involuntarily) – this is analogous to the parameter R_0 in epidemiological SIR models (Gallacher & Bright, 2021). Hoang and Lim (2016) further distinguish between "fan-out" and "propagation count" to allow for the fact that a given user may share an item of content multiple times.

Fork I

How to quantify pathways?

1. Create **recommendation maps**
2. Measure the **distance** to illegal or harmful content
3. Measure the recommender's **learning of unwanted preferences**
4. Check if the recommender displays **increasingly extreme content** to a user over time
5. Consider the **explore/exploit trade-off**

This list is not exhaustive.

Arises only in methods seeking to measure pathways as an outcome variable (Fork B)

Academics and industry researchers have expressed concern that recommender systems create pathways to illegal or harmful content. There are different ways of formalising the concept of a "pathway" in a way that can be assessed. Here, we focus on alternative formalisations where mere *exposure* to harmful or illegal content is the primary concern. We note that there are stronger forms of pathways with real world outcomes (Fork B), such as when a feedback loop between a recommender system and a user causes the user's preferences to shift over time, which in turn causes the recommender to show them more harmful or illegal content (Thorburn, Stray, & Bengani, 2022b).

Recommendation maps are graphs of recommendations made in the context of particular items of content (Brown et al., 2022; Etic Lab, 2018). For example, consider a YouTube user watching a particular video. Beside the video will be a recommender slate of "Up Next" recommendations. Clicking on each of these videos will take the user to those videos, which each have their own set of "Up Next" recommendations. We use the phrase *recommendation* map to refer to the abstract graph structure connecting each video with the videos that are recommended alongside it. Annotating a recommendation map to indicate which items of content are illegal or harmful can give a sense of how near a user is to such content – and by implication, how likely they are to encounter it – in a given context. An example output is given in Figure 2. Recommendation maps are specific to the choice of the original item of content from which the map is constructed, and if recommendations are personalised, recommendation maps will be different for different users' accounts. Recommendation maps also include all possible recommendation paths from the original item of content, of which a given user will usually only take one. User-agency will therefore play a substantial role, and this is difficult to model effectively.

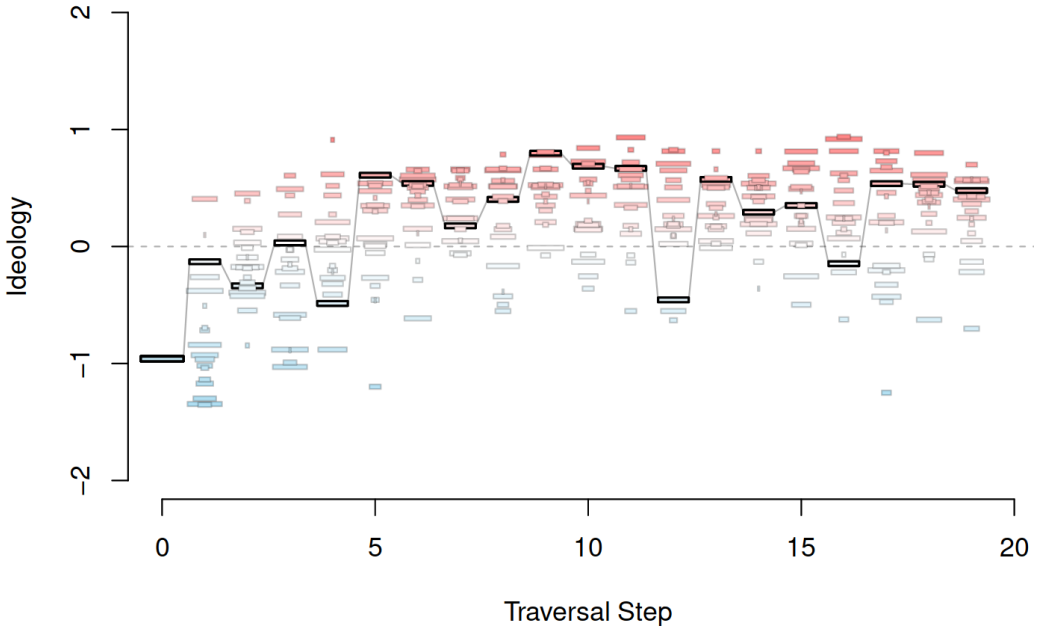

Figure 3 A visualisation of a partial recommendation map (reprinted from Brown et al. 2022).

Less comprehensively, it is possible to measure the **distance** from a user to illegal or harmful content. There are multiple potential measures of distance, including the number of clicks it would take them to reach such content and the physical time taken to reach such content. Such distance could be theoretical (i.e., the shortest path in the recommendation map, regardless of whether users take that path in practice) or empirical (i.e., how long it actually takes real users to be exposed to illegal or harmful content).

Alternately, if a recommender can **learn preferences** for illegal or harmful content and start recommending such content more frequently as a result, then this could be considered the creation of a pathway. Most recommenders are designed to be content-agnostic by default, so are in principle capable of learning preferences for any kind of content, including that which is illegal or harmful. However, it is possible to design recommenders so that they do not learn such preferences (Whittaker et al., 2021). Success at avoiding the creation of pathways of this type can be evaluated

by creating a user account which engages with illegal or harmful content, and measuring whether similar content starts to be shown to the user more frequently.

Another idea of a pathway is that a recommender might display **increasingly extreme content** over time. Pathways of this sort have been evaluated in the context of research on algorithmic rabbit holes and radicalisation (Boucher, 2022), usually by developing a way of quantifying how extreme items of content are along a relevant dimension, and then analysing whether there is a systematic bias in recommendations towards more extreme content than the user is already engaging with.

Finally, many recommender systems have one or more parameters that determine their **explore/exploit trade-off** (Barraza-Urbina, 2017). That is, the degree to which they show content to the user which they are confident the user will engage with (exploitation) versus content they are uncertain they will engage with in the hope of discovering user preferences that the recommender previously was not aware of (exploration). It could be argued that recommenders that give relatively more weight to exploration will present more diverse content to users, and hence are more likely to recommend illegal or harmful content, creating a pathway to it. Thus, quantifying the explore/exploit trade-off may provide an indirect means of measuring pathways. It is perhaps a weaker notion of a 'pathway' than those described above. Nonetheless, the fact that most recommenders will eventually show users illegal or harmful content (through "exploration") unless deliberate steps are taken to avoid this constitutes an important baseline scenario against which the existence of pathways can be measured.

5.1.3 - "Audits" of recommender systems

The word "audit" is a commonly used umbrella term in discussions about the evaluation of algorithmic systems (Costanza-Chock et al., 2022; Digital Regulation Cooperation Forum, 2022), and broadly speaking, audits do satisfy certain properties which distinguish them from non-audit methods. Typically, audits:

- Are conducted by independent third parties.
- Follow an analysis plan that is determined in advance.
- Are carried out to a high level of thoroughness or completeness.

These properties can apply regardless of the specific research questions that are being answered, and independently of the method used to answer them (so long as it is reasonably thorough or complete). For example, methods have been proposed (or, in one-off cases, trialled) for auditing recommender systems on issues such as:

- Whether "recommendations [are] influenced by … revenue" to the platform (Sandvig et al., 2014);
- "Preference-based fairness" (Do et al., 2022);
- "Algorithmic bias in job recommender[s]" (Zhang, 2021);
- "Misinformation filter bubbles" (Ramaciotti Morales & Cointet, 2021; Srba et al., 2022);
- Whether a recommender is politically biased (Huszár et al., 2022);
- Longitudinal impacts such as "polarization or segregation of information among … users" (Dash et al., 2019);
- Whether a recommender system is displaying banned content (Tracking Exposed, 2022); and
- Whether certain recommendations would have been made under a counterfactual (Akpinar et al., 2022).

To varying degrees, these methods could be directly used or adapted to assess questions related to illegal or harmful content. However, the specifics of the proposed audit methods are diverse, spanning from simply trying to upload a banned video (Tracking Exposed, 2022) to conducting large-scale platform experiments (Do et al., 2022). Rather than lumping all these methods together under the label "audits", we think it is usually more useful to talk about a specific method by situating it within the decision fork framework introduced above.

At the time of writing, there are few public examples of recommender "audits" conducted by or with the cooperation of platforms. There are occasional audit-like studies on various questions published by platform researchers (e.g. Bakshy et al., 2015), and an August 2022 news story reported that Oracle had been hired to audit TikTok's recommendation and content moderation algorithms with a focus on foreign interference (Fischer, 2022), though the methods employed and results of these audits are not publicly known. Increasingly, regulations such as the EU Digital Services Act (European Union, Article 37, 2022) and the proposed US Platform Accountability and Consumer Transparency Act are requiring that large platforms accommodate auditing processes. However, the extent to which these regulations enable direct auditing of recommender systems remains unclear; the specific methods are left unspecified in primary law (Coons et al., 2022); and existing proposals for algorithmic auditing lack standardisation (Costanza-Chock et al., 2022). That said, there are considerable ongoing efforts to ascertain which methods such audits should use, such as the Auditing of Recommender Systems Project of the German technology policy think tank SNV (Stiftung Neue Verantwortung, 2022). Moreover, with respect to the EU's Digital Services Act, the European Commission intends to publish secondary legislation on DSA compliance auditing methodologies – which might have practical relevance for the auditing of recommender systems – in late 2023.

Box 4

What makes a method good?

Below, we list some general principles for what makes an evaluation method good. Some of these may be in tension with one another and require trade-offs to be made. The evaluations below draw from the expert interviews and the literature review performed in this report.

Causal > descriptive or speculative (Fork E)
It is better for an evaluation method to identify outcomes which are caused by a recommender system, rather than outcomes which may have occurred anyway.

Standardised and commensurable > ad hoc
It is better for an evaluation method to be sufficiently standardised that its findings can be tracked over time and compared with those for other recommender systems.

Achievable > unrealistic
It is better for an evaluation method to actually be used, which means it should be affordable for a platform to perform.

Reproducible > taken on trust
It is better for an evaluation method to be able to be performed by third parties, so that the findings of the method do not need to be taken on the word of a corporate platform.

Scientifically valid and generalisable > noisy or anecdotal
It is better for an evaluation method to produce findings that are accurate and generalise across contexts (e.g., across users, time periods, types of content, etc.).

5.2 - Comparison of methods

The methodological toolkit presented in Section 4.1 contains all of the methodological elements proposed by interviewees over the course of the interviews, discussed in the discovery workshops, and gathered from the literature review. While these method elements are typically used in combination within an overall assessment method, in order to facilitate direct comparison between them, this section considers the relative strengths and weakness of each element.

Each of these methodological elements have their advantages and disadvantages, and each is suitable for use in different situations for services with different requirements. In addition, the exact combination of elements will dictate the trade-offs offered by an overall assessment method. In order to provide a comparative overview of these methodological elements when it comes to measuring the impact of recommender systems in the promotion of illegal content, we scored each element along the following dimensions: insight gained, resources required, cost, validity, and ease of standardisation. These metrics were selected as they represent the key requirements for a successful overall assessment method while also highlighting areas likely to be in tension with one another. These metrics are given in more detail in Table 1.

For each combination of methodological element and metric, members of the research team performed a round of scoring based on insights gained from both the literature review and expert interviews. Scores ranged from 1 (low) to 10 (high) for each element-metric combination. The results are presented in Figures 4–6 and the Appendix. The exact scores for each element are not intended to be insightful in isolation, but rather the overall trends and relative scores across elements should be considered.

The results from this process indicated a few key trends as well as some tensions between different methodological elements.

Table 1 Metrics used when comparing different assessment method elements.

Metric	Description	Example low score	Example high score
Insight gained	An estimate of how much new information researchers will learn about the impact of the recommender system by successfully including this methodological element. This includes insights in the wider sense, and accounts for the value of receiving null results.	Only a small degree of new information about the system will be gained.	Successful completion of this element will lead to a large amount of new insight into the impact of the system.

Metric	Description	Example low score	Example high score
Resources required	A measure of the level of resources that would be required within an organisation for them to be able to include this methodological element. This includes the skills and experience of staff members, such as data science and engineering teams, as well as physical resources, such as computer infrastructure or data. It also includes more intangible assets, such as access to willing participants or a suitably engaged customer base. This measure does not include the cash required to run a specific assessment method once the prerequisite resources are in place.	Few resources are required before this assessment method can be deployed, there are no specific data requirements and the method can be performed by non-specialist teams.	There is a substantial level of resources required for this method to be deployed successfully. This could be for highly specialist data science teams, or pre-existing hardware.
Cost	What is the expected marginal cost of performing the assessment method, assuming the required resources are available? For example, once a data science team is in place and is able to perform an analysis technique, what is the likely cost every time this method is performed?	The marginal cost of each subsequent assessment is low and does not increase over time.	The marginal cost of each assessment is high, e.g. through specific participant costs or infrastructure costs.
Validity	How well does this methodological element accurately measure the phenomenon of interest? Here, we are evaluating for both internal validity (to what extent the method measures a true effect free from spurious influence) and external validity (how applicable the findings are to other contexts outside that immediately studied).	The results may only apply to the specific context in which they are studied, or the results are driven by spurious or confounding factors.	The element measures a true real-world effect and applies well to contexts outside the specific scenario.
Ease of standardisation	How easy is it to standardise this methodological element so that it can be repeated over time or across sectors and can be applied to other platforms? This includes how unique the research skills required to perform this approach are, as well as the data requirement or access to proprietary hardware/software.	Highly specific and unique approach, possibly new or immature, which cannot easily be applied to different contexts.	A mature methodological element that can be repeated elsewhere, allowing for comparisons.

Trade-offs between insight and required resources

We find a clear positive relationship between the resources a service is required to have in order to use a methodological element and the insight that this method will provide (Figure 4). This trade-off is perhaps unsurprising, and points to a clear conclusion: there exists no single overall assessment approach which is both low resource and high insight.

Experiments conducted via A/B tests scored the most favourably across the combination of these two metrics (Figure 4, E1), generating a high level of insight for a comparably lower level of resources required. This is driven by the fact that conducting simple A/B testing via a split-test is relatively easy for a platform to do, and in this simple configuration does not require much in the way of specialist infrastructure or statistical expertise. As the complexity of the experimental design increases, both the infrastructure and data science requirements increase to the point where running large multi-armed bandit type experiments could have a very considerable marginal cost attached for each new experiment. Given this, it suggests that the simpler style of A/B tests which compare two versions of a recommender systems, each with different design choices, within a small but representative segment of the platforms user-base, represents a good trade-off between insight and the required resources.

It should also be noted that on-platform experimental approaches need to be combined with an additional measurement approach (D1-4) to record the outcome of interest and this will increase the resources required for this approach. Some of these approaches will require only a small degree of resources, such as user-reporting or user-surveys, while others, such as the creation of sock-puppet accounts and post-hoc analysis would be far more resource intensive.

A similar tension is demonstrated in the case of off-platform experiments such as the development of browser-extensions for collection of user experience and behaviour data. As these extensions collect data in a naturalistic setting, they aid ecological validity and gives a good degree of insight. However, the cost of developing a custom browser extension is often extremely high, both in terms of the initial development and the ongoing operation (Kohlbrenner et al., 2022). Additionally, given the range of platforms, web browsers and user systems, each creating a standardised approach to developing this type of experiment is challenging.

Standardisation helps lower costs

Finally, we find that there is a negative relationship between the marginal cost of including a methodological element within an assessment approach (assuming an organisation has the capability required to perform this assessment) and the ease of standardisation of this element (Figure 6). This implies that low-cost high-standardisation elements are particularly good approaches for consistent deployment over time or as a minimal threshold to set across services and industries. These low-cost approaches will not give the greatest insight, however (Figure 4). An example of this trade-off is given by platforms which rely on user-reports as the key metric to evaluate changes to their systems. This element scores highly for standardisation and has a very low marginal cost, however, the insight generated is among the lowest across all elements. This means that while it could be a good initial assessment approach, it is unlikely to be sufficient in the longer term for larger platforms with access to more resources. The exception to this trend is civil society reports, which while relatively cheap from a platform perspective, are difficult to standardise due to the sheer number of ways that this could be done.

When considering the marginal cost of an element it is important to highlight that there may be some clarity needed on who should pay for these assessments to be carried out. If the cost is carried by the platforms how this funding is both secured and distributed across teams is an open question, which will likely depend on the resources available to a service.

Overall, this comparative approach indicates that there are a number of trade-offs that must be considered when selecting methodological elements from which to build an overall assessment method. The overall scores for an assessment will be determined by the contributions from each element selected at each fork in the methodological toolkit.

Looking at the specific elements, while there is no single element which scored perfectly across all metrics, **we find there are a selection of elements which present a positive trade-off across the range of metrics. These elements include on-platform experimental approaches, such as A/B tests, and recommender system debugging, as well as the more sophisticated quantitative approaches focused on the analysis of observational data, such as epidemiological models and causal inference on observational data.** Conversely, this comparison suggests that experimental approaches using sock-puppet accounts, off-platform 'lab' experiments, whole platform simulations, and ad-hoc survey approaches all score comparatively poorly.

Figure 4 Comparison of resources required to use a methodological element against the insight it is likely to generate. Letter-number references are given to Table 3. See appendix for legend.

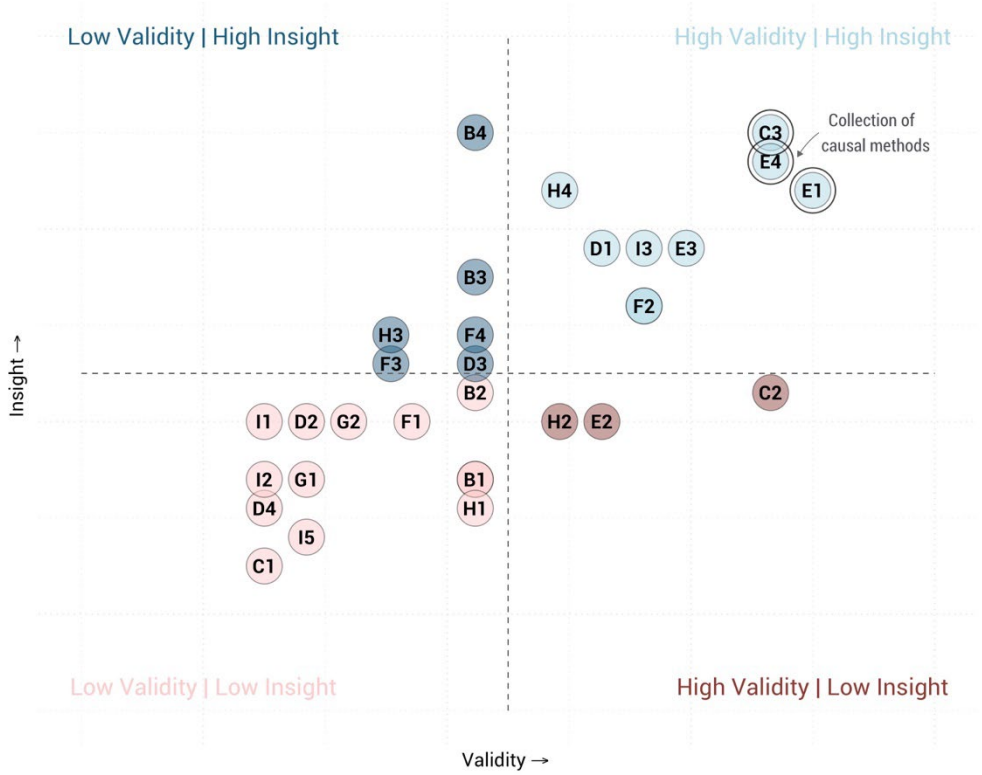

Figure 5 Comparison of validity (internal and external) of a methodological element and the insight it is likely to generate. Letter-number references are given in Table 3.

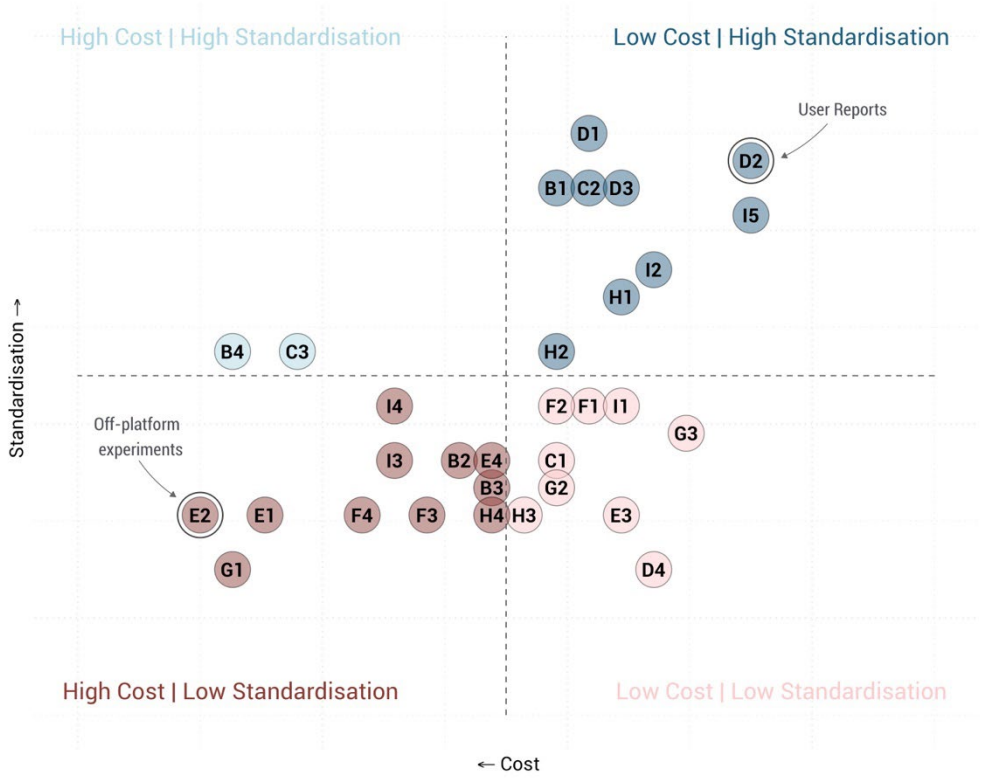

Figure 6 Comparison of the ease of standardising each methodological element and the marginal cost of each subsequent deployment. Letter-number references are given in Table 3.

5.3 - Platform guidance

Our findings in this report indicate that there is no single best method and no "magic formula" when it comes to evaluating recommender systems. Instead, each method has strengths and weaknesses, and different methodological elements work together to produce an overall approach. Similarly, it is advisable to combine different overall evaluation methods, as no single approach can capture the complete phenomena of interest.

Additionally, there is an important distinction between **formative** and **summative** evaluation. Platforms should undertake some forms of evaluation frequently or continuously to inform the ongoing design of their recommender product (formative evaluation) and monitor for situations where external factors alter the way their recommender system interacts with illegal or harmful content (summative evaluation). For context, we also suggest that regulators focus their efforts on articulating best practice for formative evaluation and responsible development – not just summative evaluation of a "static" algorithm – and to consider what would constitute negligence or reasonable due diligence in the process of developing a recommender system.

The combination of methodological elements that are adopted for evaluation will vary depending on the service size, resources, and the type of service that they provide. For example, for a small service it might be advisable to collect user feedback by providing a reporting option for problematic content and engage with users on a semi-regular basis to continuously collect data, and then perform a functional test / stress test (G3) before rolling out a large-scale change to the recommender system. These elements may present a positive trade-off between costs and insight. For a larger service they might also opt to additionally conduct a more formal audit, potentially performed by an independent organisation, of the entire system at fixed intervals (e.g., every year)

to give more insight (see 4.1.3). However, there are few industry-wide norms for how to perform these more comprehensive audits. One facet which spans multiple services regardless of size is that it is good practice to focus on the robustness and repeatability of a method so that it can be used over time and any changes can be monitored, rather than a single, very expensive approach that only provides a snapshot at a single point in time.

Based on the evidence gathered for this report, there are a number of principles that facilitate the effective evaluation of recommender systems. The following subsections summarise these findings.

Box 5

Best practice for different kinds of service

The methodological toolkit and comparison of methods given in Sections 4.1 and 4.2 are intended to provide service providers with a guide for designing the best assessment approaches for the recommender systems in use on their platforms. Here we provide three quick examples of what this looks like for platforms of varying sizes and sectors.

A large video sharing service

Large user-to-user platforms, such as video sharing services, contain high volumes of ever-changing media content and make heavy use of recommender systems to help users find content they may be interested in. These platforms also typically have many resources available and extensive development and data science teams in-house.

As such, research indicates that platforms may benefit from building a robust and ongoing evaluation strategy which takes into account multiple data sources for flagging and detecting illegal and harmful content, and uses a combination of data science (E3) and A/B tests (E1) to understand the impact of all significant changes to design of their recommender systems on the spread of such content, before widespread deployment.

Research also indicates that platforms could benefit from being aware of the risks that occur due to the posting of links that re-direct users to off-platform content which may be illegal or harmful. To help mitigate this, these platforms should invest in collaborative efforts with civil society organisations academia, and the wider industry.

A medium-sized dating app

Medium sized platforms, especially those with a focus on user discovery such as dating apps, also make substantial use of recommender systems, although these are often less prominent. These platforms are often less studied, and so the potential harms occurring from recommender systems are less well understood. Their users are likely to be fairly well engaged with the platform and service and more likely to actively participate in research.

By refocusing their efforts on conducting in-depth user surveys (F1) and interviews (F3, F4) and combining with observational data analysis (C2), platforms can may be able to better understand and uncover the risks of their recommender systems.

A small specialist news aggregator

Small platforms make up the long tail of service providers online. These platforms each individually host a small amount of content, but are common and so collectively make up a large proportion of online content. These platforms are often under-resourced, commonly with only one or two full-time staff and ad hoc software development contractors.

These smaller, less resourced, platforms may benefit from using off-the-shelf and commercial tools (D1) to assist with the detection of harmful content that is hosted on their platforms and exclude any content from their recommender systems if it scored moderately highly by these systems. If ongoing monitoring using these tools is too expensive, they can also be used to perform a functional test / stress-test for major changes to recommender systems prior to deployment (G3).

Plan ahead, know what to measure, and measure it at the right time

Model evaluation should ideally begin before adding new features or making changes to a recommender system on a platform (Ind5, DW3). As one interviewee noted (CS4), it is safer and cheaper to check an algorithm before it goes live than when it is already in production.

The exact timing of the evaluation of a recommender system depends on a number of factors, including the size and scale of the platform, the risk of harm occurring, the type of content the platform offers, the audience, and the role the recommender system plays in determining the content which users interact with. For a trusted platform such as a public broadcaster, this might mean doing highly intensive functional testing / stress-testing before any change to a recommender system is rolled out, as their tolerance for an inappropriate recommendation is likely to be very low. For a larger service, this barrier is likely to be too high as they are deploying changes to their recommender systems much more frequently (possibly hundreds of times a day) and their risk tolerance is also higher. For such a service it might be more appropriate to perform evaluations using real-time feedback from users, and more active testing at a fixed interval when a more experimental recommender design is to be deployed.

Evaluations can also be done in a graded and ongoing fashion. For example, when deploying significant changes to a system, it was reported as good practice to deploy this in a gradual manner across user-groups and collect continuous feedback on the impact of this change, only proceeding if this feedback is, on balance, positive (Ind5). This process will increase the requirement for statistical rigour in the analysis, and so should be deliberately planned in advance. In practice this might look like initially testing a change on a group of highly engaged and trusted users (such as paid testers or even organisation staff) and collecting qualitative feedback via user surveys and interviews, before rolling out progressively over the user-base and collecting quantitative feedback via user reporting and other business metrics.

Identifying examples of when illegal content has been recommended to users, and working backwards from these examples (recommender system debugging, E4), could be an effective way of evaluating recommender systems (DW2). However, in order to conduct this kind of post hoc analysis, platforms have to systematically collect data on what content is suggested to users and document why certain users see that content. Similarly, if a platform does not log what content a user actually sees (as opposed to content they could have seen if they were on the platform at that time) then it is difficult to use some of these measures effectively. However, this raises privacy concerns both around data protection and user rights to privacy more broadly.

The use of model cards within machine learning development demonstrates a similar example of the benefit of forward planning (Mitchell et al., 2019). Model cards provide a simple checklist of requirements for a developer to fill in prior to deploying a model which give brief descriptions of the intended use cases of the model, along with a description of the data used for training and evaluation, the intended use-case, any out-of-scope uses to be avoided, and any ethical

considerations. Evidence from industry experts suggests that this type of written exercise helps anticipate unintended consequences, and provides a more accountable development process (DW1). The use of model cards also makes any post hoc investigation much easier. In larger organisations, this allows non-technical teams to be involved at an early stage in the process of evaluating machine learning models. The creation of model cards is not a resource-intensive method, but companies need to use it deliberately and consistently, and this is true regardless of whether the model cards are intended for internal use or public publication. The evidence gathered in this study indicates that model cards are widely used across the industry, and are quickly becoming best practice.

Incorporate input from users (cautiously)

Assessing recommender systems is intertwined with the assessment of other elements of content management on a large platform. Recommender systems do not work in isolation but use signals generated by content moderation teams and internal security teams in order to identify which content is suitable for recommendation, which content should be downranked, and which content should be removed from the platform all together (see Box 3). It is also important that content that has been flagged or rated low/negative[6] by users is evaluated, and that this information can be used to retroactively improve the system to avoid similar mistakes (E4). This improvement could come from engineering efforts to address content processing or system-interaction limitations, or it could be used to re-train the recommender models themselves by removing the problematic content from the training set, or even using these removals as signals of the content not to recommend in a reinforcement learning paradigm (Afsar et al., 2022).

The level of content flagged by users as potentially harmful can also be used as a key metric for evaluating changes to a system (Ind5). User reporting of problematic content is a relatively simple and inexpensive solution to detecting the false negatives[7] from content moderation, however it is reliant on having a suitably engaged user base, and the metric does lag behind the problematic content being shown to users. Although it leads to limited insights when used as an evaluation method for an entire recommender system, user reporting still plays an important role in the ecosystem of tools for safely running a recommender system.

It is important that this measure is robust, however. Systems designed to allow users to rate content negatively "can be gamed" (CS3), raising the risk that a malign actor might try to get legitimate content removed or high-profile users banned from a platform. As a result, single bursts of user reports are not a reliable enough metric to use in isolation. Instead, a more reliable metric is the collection of diverse low/negative signals. These diverse negative signals will, for example, originate from multiple users, in multiple regions, with varying account types (creation dates, user profiles, online behaviours, etc.) over a longer time period. This diversity is difficult to imitate due to its inherent randomness.

Twitter's Community Notes programme (previously known as Birdwatch) was highlighted as a potentially promising implementation of this (CS3). The platform allows Twitter users to add notes

[6] In software engineering, low or negative user ratings of content on U2U services can be programmed to send "negative feedback/signals" for the recommender system; negative signals can be used to indicate that the content is of low quality, which in turn would be deprioritised by the recommender system.

[7] False negatives in this context refers to when illegal (or violative) content is misclassified as innocuous by a services content moderation process. This has a variety of causes, including a miscalculation by an automated classifier or ill judgement by a human moderator.

to tweets that are misleading or otherwise problematic, along with the reasons why, and was cited as an example of encouraging users to provide feedback on content. However, such systems could also be gamed by bad actors if no safeguards are implemented. The actual impact of this programme on reducing harm is however unknown. Facebook has gone to great lengths to find ways to allow users to provide negative feedback on the content presented, making it easy to report a post for a specific reason by simply selecting the problem category or flagging the content. This type of encouragement of nuanced feedback from diverse users can provide excellent signals for use in wider systems.

Invite additional scrutiny

If platforms invite additional scrutiny of their assessment methods, then this has the potential to improve both the internal validity of these methods, by checking for accurate implementation and analysis, and external validity, by checking for appropriate methods selection and completeness. Additional inspection of the results of any assessment carried out can also help to ensure these results are interpreted appropriately. This view is reflected in a recent review of algorithmic audit processes (Meßmer & Degeling, 2023) where they advise that services set up an assessment process which brings together various elements, issues, stakeholders and assessment types in order to make a well-rounded assessment of the impact of their systems.

This scrutiny can be conducted either by internal teams not involved directly in the assessments, or by external experts from academia and civil society, depending on the level of resources and expertise available 'in-house' and the appetite for engaging with external teams. For some platforms this type of review is already common practice. One platform (Ind5) explained how they have an internal review process of any proposed change to a system whereby a committee will form and evaluate the various assessment results and decide whether this change should go ahead. This committee is comprised of members from different departments, including product design teams, engineering teams, trust and safety teams, and the different competing interests will be evaluated. This internal transparency of the assessment process coupled with empowering the committee to make binding decisions on new features was argued to help build a more robust and reliable product and mitigate against harm.

Share tools

There are several examples of open-source tools being developed which can help with certain elements of the assessment process. For example, Meta recently released a safety tool that can identify copies of images and videos that may violate community standards and can be used by platforms to prevent the distribution of content related to terrorism and extreme violence (Clegg, 2022). Similar efforts should be developed to share the tooling that larger platforms use internally for assessing the impact of recommender system design choices, and making these tools available to the wider community. These tools are likely to be expensive to develop, and therefore only the larger platforms will have access to the required level of resources. By publishing these tools open-source, developers and researchers outside the platform can help make those methods more robust, therefore benefiting the original organisation. In addition to large platforms, civil society groups can also open-source the tools they develop. A good example of this is the browser-extension (E2) software developed by Tracking Exposed which can be used for investigating the recommender systems used by platforms including YouTube and TikTok (Tracking Exposed, 2023). These open-source collaborative tools have been used by researchers and journalists outside the organisation.

Adopt best practice from machine learning operations

Machine learning operations (MLOps) are a set of modern best practices and techniques for managing the end-to-end lifecycle of machine learning models (Microsoft, 2023). These best practices describe the use of automated tools and processes to manage the development, testing, deployment and monitoring of machine learning models. In practice this may include techniques such as version control for model development, version control for datasets, automated testing, and validation of models prior to deployment, and ongoing monitoring of the performance of deployed models, with defined processes for error analysis.

MLOps is a burgeoning and fast-moving area of machine learning (do Prado, 2023) and so we do not prescribe any specific tools or techniques here, but rather than the principles of MLOps should be baked into the development and deployment of recommender systems. These best practices are quickly becoming standard across industry, often with smaller, more agile companies leading the way. The UK government has provided guidance in this area, initially from the National Cyber Security Centre on safe development for machine learning (NCSC, 2022), and secondly in the form of tooling in GCHQ's Bailo machine learning development and deployment platform (GCHQ, 2023).

5.4 - Adoption of methodologies for assessing recommender systems

Services can apply a wide range of assessment methods depending on the various factors of interest for the assessment of a recommender system. During this research three main groups of methodological elements were identified, as discussed in Section 4.2. Each element provides insight into different aspects of recommender systems, from testing the actual recommender system through to examining user perceptions and real world outcomes. A combination of various methodological elements can thus reveal more information about the impact of recommender systems on societies. However, adopting some of the methodological elements also brings several challenges for digital services. In this section of the report, we evaluate the possible barriers preventing services from adopting evaluation methods, then some potential consequences of them doing so.

5.4.1 - Barriers

The adoption of specific methodological elements can create certain challenges for services. Due to the differences between services in their size, purpose, design or revenues, these challenges may be different for different services. Big tech companies are better equipped to conduct evaluations, while for smaller services the assessment methods can pose much bigger challenges. Large companies already conduct internal audits of their recommender systems and hence have the necessary know-how and resources. "With smaller platforms, you need to have a different set of expectations of what they can do" (CS5). As the challenges are different, thus, the solutions may also differ from service to service.

Several factors influence whether platforms can effectively adopt measures to evaluate recommender systems:

Monetary costs

One of the most significant barriers to the adoption of assessment methods is cost and resources. As mentioned in Section 4.2, assessments which bring better insights can be costly. Such methods require skilled staff, time, capacity, and robust systems, which can significantly impact the overall costs. The low-cost assessment strategies, such as self-reporting, do not require a large number of

staff or special technical skills and rely on the willingness of users to participate. This group of methods can thus be easily applied across a wide range of services. Another element which lies on the cheaper end of the spectrum but can provide better insights is recommender system debugging. However, it requires the substantial ability to work with data, and thus platforms will need to bring staff with relevant skills and tools able to process it.

Exactly determining the cost of an assessment method is difficult however. One interviewee – an expert from the tech industry – observed that "the resources required to conduct comprehensive and ongoing assessment of a system with multiple layers of recommendation and interaction with other systems is huge" (CS4). While the largest services companies may be able to cover these costs, for smaller services such as start-ups, this can be a material burden for their business. Nevertheless, as mentioned by CS1, methods like quantitative assessment of "observational" data are more expensive when adopting them for the first time, as it is necessary to set up the infrastructure. As evaluation should be conducted continuously, the ongoing cost will reduce once the system is running properly.

In addition, indirect costs may occur from changes to user engagement following recommender system assessments. An assessment may identify an issue related to safety which can only be resolved by making changes that lower user engagement, and thus service revenue. An additional scenario which may impact service costs is when a proposed update to a recommender system, designed to increase engagement, is delayed while the service assesses the impact of changes on safety metrics. This delay increases costs to the service, compared to if they had introduced the revised recommender system across their service at an earlier stage.

In a survey of 54 AI practitioners' attitudes to the implementation of ethical AI in industry, they were asked to list challenges of developing AI products ethically. The most common answer, chosen by 33% of respondents, was that it would "incur additional costs" (Morley et al., 2021). Indeed, an intervention in a recommender system that changes the outcome metric from solely engagement-focused metrics to some other less sticky[8] goal would require significant investments from technology companies, such as an interdisciplinary team dedicated to understanding sociotechnical problems in recommender systems (Beattie et al., 2022). It is likely that engagement will remain a significant outcome measure however given that it provides a useful signal about what is valuable to the user (Bengani et al., 2022). Moreover, a decrease in engagement may be costly, especially as the business model for many platforms relies on ads. These costs are, however, crucial to protect users in the sector given the high potential of risks the users are facing. As platforms are increasingly subject to statutory duties relating to the protection of users, the cost may be seen as necessary to run business in this sector.

Reputation

In multiple interviews for this research, interviewees stressed that the main motivation for services is profit. Therefore, any reputational damage that results from revealing the flaws of their recommender system can be another barrier to services evaluating these systems. This can particularly be an issue when it comes to illegal content. If services discover as a result of the assessment that their systems are recommending illegal content, they may face enforcement action under new online safety regimes such as the OSB and DSA. A possible solution might be to ensure

[8] Stickiness is a measure of user retention; specifically, how often users are returning to a platform and the amount of time they spend on the platform. Sticky goals are those oriented towards user retention.

that the results of recommender system assessments are confidential, or released only to the regulator, as such a measure may decrease the risks connected to reputation and so platforms may be more motivated to perform assessments. On the other hand, keeping assessments confidential will decrease the possibility of external audits by civil society or governmental institutions as a control mechanism which are (according to several interviews) necessary to keep platforms accountable.

Research ethics

While some methodological elements, such as surveys or interviews with users, belong to standard methods in academic research where ethical concerns are typical minimal, other approaches, particularly experimental approaches, can raise issues of research ethics. Assessment methods which include experiments with actual users of a service pose a risk to ethical standards, especially when there is a risk that users may end up being exposed to illegal content as a result. For ongoing on-platform experiments such as A/B testing, users may consent to participating in this research when they create an account on an online service. While it is desirable for users to be informed about individual experiments, it may not be practical due to their frequency and the risk of influencing user behaviour to the effect of biased outcomes. Ideally, such methods would undergo some form of research ethics review, preferably by a third-party organisation, to ensure that the assessment is conducted ethically.

Privacy and data protection

Some assessment methods, particularly the observational methods, may require the use of personal data. Services must be assured that they can do so in a way that adheres to data protection regulation, and that they have the necessary data governance protocols in place to protect the privacy of their users (e.g. through anonymising that data). Services must also ensure that any personal data they collect and process is not used for purposes other than the one originally intended, unless there is a lawful basis for doing so. These issues also apply where a service outsources the assessment of its recommender system to a third party. In such cases, the service would need to be assured that the third party is collecting and processing any personal data in accordance with data protection regulation and in line with ethical standards more broadly. In some circumstances, services – particularly the smallest – may feel unable to adhere to these obligations, meaning they would not be able to proceed with an assessment of their recommender system (or would need to choose an alternative method that poses fewer data protection risks).

Lack of available standards

Standards bodies have begun to develop several standards that aim to support industry in scrutinising their use of algorithmic systems. This includes for example a proposed standard from the ISO that would “provide guidance for organisations performing AI system impact assessments for individuals and societies that can be affected by an AI system and its intended and foreseeable applications” (ISO AWI 42005). However, there are few standards that relate specifically to the assessment of recommender systems. This in turn makes it more difficult for services to understand which assessment method to deploy and how exactly that method should be performed (n.b this paper does not intend to go into that level of detail). Moreover, without these standards, it is

challenging to foster common practice across industry, which means that the results of any assessments are difficult to compare and contrast from one context to the next.

5.4.2 - Unintended consequences

Risk of feedback loops

It is important to be aware of the risks of unintended consequences when changing how recommender systems are assessed. A particular risk is that a service creates new metrics as part of those assessments, which go on to create perverse incentives for users and those seeking to game the recommender system.

An example of this is the introduction of Meaningful Social Interactions (MSI) metrics by Meta on Facebook, which was introduced in 2018 as the main business metric used to rank content in its News Feed. MSI was calculated from the weights of predictions for different types of reactions and gave greater prominence within the recommender system to posts with long, extensive comments in an attempt to capture positive and engaged conversation (Cameron et al., 2022; Facebook Papers, 2019, 2020; Metz, 2021). Reports have subsequently suggested, however, that MSI resulted in an increase in negative interactions (Facebook Papers, 2023; Leqi & Dean, 2022), as the model learned to prioritise sensational content (Meserole, 2022). It has also been suggested that change also appeared to lead to the creation of further harmful content in addition to the promotion of existing harmful content (Leqi & Dean, 2022). Facebook has since made considerable changes to MSI and have removed some factors of engagement based ranking. , For example, for political content, Facebook have removed the reinforcing signals that its recommender system uses to rank a post based on the prediction that a user will share it or comment on it (Iyer, 2023). As a result, the platform received qualitative feedback from users that their feeds were better in quality and worth their time due to a reduction in political content more "worth your time" as well as measurable reductions in quantitative metrics indicative of negative online interactions (e.g., anger reactions on Facebook posts) (Horwitz et al 2023).

Box 6

Case study: Assessing recommender systems and terrorist and violent extremist content

Assessing the role of recommender systems in promoting terrorist and violent extremist content (TVEC) poses specific challenges. Interviews with experts in online terrorism and counterterrorism revealed that these challenges stem primarily from the relationship between user agency and the promotion of content by recommender systems, the interaction between social media platforms and off-platform content hosting sites, and the difficulties in setting appropriate thresholds for content definitions (A1, A2, Ind2, CS5). The establishment of a common cross industry definition of TVEC also presents a major challenge[9].

User agency

While recommender systems can lead to an increase in exposure to TVEC, this effect is most strongly focused on those individuals who already hold extreme beliefs, and who are likely to also actively search out this content. This was expressed by some experts as the key dilemma in

[9] GIFCT-TaxonomyReport-2021.pdf

measuring the effect of recommender systems on processes of extremism (Ribeiro et al., 2023). For a behavioural change effect to occur, exposure must be matched with vulnerability (A1, A2). Assessment methods must therefore consider whether the design choices within a specific recommender system played a role in increasing access to harmful content which the individuals are already actively searching out and should monitor users over extended periods of time to measure changes in user preferences or attitudes.

Off-platform content

Off-platform content is another challenge, as the interviewed experts advised that very little TVEC that amounts to relevant offences is hosted on mainstream platforms (Ind2; Tech Against Terrorism, 2021). This is due to the success of collaborative efforts such as the GIFCT hash-sharing database and Tech Against Terrorism's TCAP, as well as internal detection methods used by social media platforms. As a result, any assessment method investigating recommender systems and known TVEC on mainstream platforms will be unlikely to find any direct relationship. However, the larger concern raised was in the promotion of URLs that link to content hosted elsewhere, away from the social media platform where the users are located (Ind2). In these cases, the URLs shared on the initial platform may still be intended to show users potentially illegal content, but as the material is not uploaded to the original platform, it may not be caught by moderation systems or hash-matching systems. This is an effect which was observed following the 2022 Buffalo attack, where URLs were posted on larger platforms which linked back to smaller, less well-known platforms hosting footage of the attack (Ofcom, 2022). Whether mainstream platforms should follow all URLs shared on their service in order to remove those that link to TVEC is an open question.

Similarly, it was highlighted that many extremist and terrorist groups have moved their online operations to less moderated social spaces, and use mainstream platforms as a way of recruiting new members through the sharing of less harmful content (A1, Ind2). Assessment approaches could therefore also consider whether users end up leaving the platform, either definitively or for short periods, due to content which has been promoted to them, even if this content is not itself TVEC.

Content decisions

Finally, while TVEC may not always amount to illegal content, there is a considerable amount of sub-threshold content that can still cause harm (A2). Measuring the long-term consumption of sub-threshold material is key for any assessment approach. There is often a lag between the creation of TVEC connected with fast-moving events and when it is removed from a platform and added to a hash database (CS5). Assessment methods should focus on the spread patterns of TVEC and how they differ from those of non-harmful material to understand whether the design choices of recommender systems increase or decrease the consumption of this content during the viral phase.

6 Recommendations and areas for future research

As discussed in this report, the recommender systems used within U2U services are often highly complex and involve multiple interacting components. This complexity requires a nuanced approach to any assessment, which takes into account the specific design choices and platform-specific affordances that shape these systems. While there has recently been significant progress in

developing assessment methodologies which capture this complexity, there are a number of areas that should be considered in future research.

Consider the systemic risks to online harm that recommender systems may present, beyond the promotion of illegal content, and how to evaluate for these

Recommender systems can lead to new and unique forms of online harm, beyond the promotion of harmful or illegal content, through the inappropriate combination of content which would not be harmful on its own. Examples discussed over the course of this project included the promotion of eating disorder content following a news article about famine, or pairings between content discussing terror attacks in a certain location and “what's on locally” articles for the same location. These are not “failures" of the recommender system, as it is clear why this type of content might be linked together, but the wider societal context is missing and is what might lead to the harm for users. A related example is given by a NATO report (Fredheim et al., 2019) which documented how the "suggested friends" feature on Facebook was revealing sensitive information on formations of military groups. This is unlikely to constitute illegal content, nor is this information harmful on its own, but when revealed in a specific situation, such as during a conflict, it could lead to substantial harm.

In this work we have focussed primarily on the role of recommender systems in promoting illegal and harmful content, however a wider consideration of how evaluation methods can capture the risks of this wider type of systemic harm should also be considered. Developing suitable assessment methods for this will likely include many of the methodological elements we have discussed so far, but will also require additional steps. Meßmer & Degeling (2023) propose a potential solution to this through a risk-scenario-based audit process. In this proposal, the initial stage of any audit consists of a meeting of diverse stakeholders from within a service, as well as their users, civil society organisations, legal experts, and researchers who will collectively define and prioritise risk scenarios to capture those risks which might otherwise not be considered.

Ongoing assessment and responsible product development

In this report we have discussed the importance of moving from a "product" frame to a "process" frame when it comes to assessment approaches. The same holds true for regulators. Rather than focusing on "best practice" regarding specific designs, attention should shift toward "best practice" regarding the process of responsible product development. This approach requires methods to be viewed as formative rather than summative assessments, and appropriate evaluations should be performed when changes are made to the algorithm, starting well in advance of these changes being made.

Many areas of this shift require further research to be successful. A key consideration is determining how often evaluations should occur in practice and what constitutes a "major" or "minor" change to a system. In addition, it is important to consider models of industry consultation that do not require formal, institutional meetings with regulatory bodies, but in which a multistakeholder approach is still retained. This may involve collaborating with external experts from academia or civil society who have experience working with social media data and platforms, and who could leverage these connections to conduct appropriate research which is shared with regulators. This approach would provide an alternative means for regulators to engage with industry and benefit from insights and expertise outside of formal institutional structures. The potential benefits of this approach should be explored further in future research.

Content moderation and recommender system interaction

Recommender systems and content moderation systems are both crucial components of U2U platforms, determining which content is surfaced and what is removed. These systems are closely entangled and influence each other, at times exacerbating harm. For example, recommender systems may surface violative or borderline content that content moderation systems failed to flag, while also influencing the content moderation systems by biasing the human reporting towards certain types of content. This interplay facilitates harmful content to be promoted and distributed to users. As such, further research is needed to better understand the dynamic between these systems and their impact on user experiences.

Support for promising initiatives in this space and need for further collaboration

In response to the view that online services industry lacks the auditing, evaluation, and accountability mechanisms that are common in other industries (Raji et al., 2022) there is a growing collection of civil society projects which look to develop tools for effective algorithmic assessment. These initiatives should be supported, and efforts put in to further the number of stakeholders involved with these projects and access to the required resources. Examples of such initiatives include the Open-Source Audit Tooling (OAT) Project from Mozilla[10] which is focused on developing the open-source tooling, methodologies and resources required to support algorithmic auditors. This includes developing an online community, the 'Algo Audit Network', which currently has over 400 participants on Slack.

More formal multistakeholder collaborations such as the Global Internet Forum to Counter Terrorism (GIFCT), the Christchurch Call (CCU), and Tech against Terrorism's Terrorist Content Analytics Platform (TCAP) have had a measurable impact on reducing the sharing of TVEC content across mainstream social media platforms. This model of collaboration should be replicated to address wider online harms, including those arising specifically from recommender systems, as well as other forms of illegal content such as CSAM or the trade of illegal products online.

Finally, in order to support third-party initiatives and multistakeholder collaborations a standardised way of communicating the results of an assessment should be developed (Meßmer & Degeling (2023)) as this will allow for easier cross-platform comparison, monitoring of trends over time, and the sharing of best-practice across industries.

[10] https://foundation.mozilla.org/en/what-we-fund/fellowships/oat/

7 Conclusion

Understanding the impacts of recommender systems is challenging and requires that different methodological elements be combined to form an overall assessment method. Applying these varied methods is essential to understanding the risk that users are being exposed to illegal or harmful content as a result of these systems. In some cases, technology firms already use these assessment methods, but only do so internally and in ad hoc ways, and as such it is impossible to evaluate, from the outside, the efficiency of their assessments.

Having evaluated a wide range of methods, we can conclude that the most effective, most efficient, assessments will involve multiple elements incorporating observation, experimentation and user self-reporting. Indeed, evaluating the relative strengths, weaknesses and trade-offs of these methods reveals that the particular combination of assessment methods will vary by type of recommender system. However, we have identified a set of elements that demonstrate a good balance across a range important considerations. These elements include forward looking on-platform experimental approaches, such as A/B tests, and backwards looking investigative approaches such as debugging. In addition, sophisticated quantitative approaches that focus on the rigorous analysis of observational data can provide valuable insight. With recommender systems evolving rapidly—both through user input and design changes by engineers—the assessment methods themselves need to be diligently evaluated and improved over time.

Business priorities need not be compromised by rigorous evaluation exercises. Indeed, by committing to assess and scrutinise the impact of their recommender systems, online services can enhance their reputations, reassure their investors, and ultimately provide a safer and more enjoyable experience for their users.

8 References

5Rights Foundation. (2021). *Pathways: How digital design puts children at risk*. https://www.revealingreality.co.uk/2021/07/20/report-launch-pathways-how-digital-design-puts-children-at-risk/

Afsar, M. M., Crump, T., & Far, B. (2022). *Reinforcement learning based recommender systems: A survey* (arXiv:2101.06286). arXiv. http://arxiv.org/abs/2101.06286

Aggarwal, C. C. (2016). *Recommender Systems*. Springer International Publishing. https://doi.org/10.1007/978-3-319-29659-3

Ahlgren, J., Berezin, M. E., Bojarczuk, K., Dulskyte, E., Dvortsova, I., George, J., Gucevska, N., Harman, M., Lämmel, R., Meijer, E., Sapora, S., & Spahr-Summers, J. (2020). WES: Agent-based User Interaction Simulation on Real Infrastructure. *Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering Workshops*, 276–284. https://doi.org/10.1145/3387940.3392089

Akpinar, N.-J., Leqi, L., Hadfield-Menell, D., & Lipton, Z. (2022). *Counterfactual Metrics for Auditing Black-Box Recommender Systems for Ethical Concerns*. https://responsibledecisionmaking.github.io/assets/pdf/papers/24.pdf

Andrus, M., Spitzer, E., Brown, J., & Xiang, A. (2021). *'What We Can't Measure, We Can't Understand': Challenges to Demographic Data Procurement in the Pursuit of Fairness* (arXiv:2011.02282). arXiv. http://arxiv.org/abs/2011.02282

Araujo, T., Ausloos, J., Atteveldt, W. van, Loecherbach, F., Moeller, J., Ohme, J., Trilling, D., Velde, B. van de, Vreese, C. de, & Welbers, K. (2022). OSD2F: An Open-Source Data Donation Framework. *Computational Communication Research*, *4*(2), 372–387. https://doi.org/10.5117/CCR2022.2.001.ARAU

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132. https://doi.org/10.1126/science.aaa1160

Barraza-Urbina, A. (2017). The Exploration-Exploitation Trade-off in Interactive Recommender Systems. *Proceedings of the Eleventh ACM Conference on Recommender Systems*, 431–435. https://doi.org/10.1145/3109859.3109866

Beattie, L., Taber, D., & Cramer, H. (2022). *Challenges in Translating Research cto Pratice for Evaluating Fairness and Bias in Recommendation Systems*. 528–530.

Bengani, P., Stray, J., & Thorburn, L. (2022, April 27). What's Right and What's Wrong with Optimizing for Engagement. *Understanding Recommenders*. https://medium.com/understanding-recommenders/whats-right-and-what-s-wrong-with-optimizing-for-engagement-5abaac021851

Berger, J., & Milkman, K. L. (2012). What Makes Online Content Viral? *Journal of Marketing Research*, *49*(2), 192–205. https://doi.org/10.1509/jmr.10.0353

Beutel, A., Chi, E. H., Diaz, F., & Burke, R. (2020). *Responsible Recommendation and Search Systems*. FAccTRec, New York, New York.

Birtwistle, M. (2021, December 17). *Addressing barriers to responsible innovation—Centre for Data Ethics and Innovation Blog*. https://cdei.blog.gov.uk/2021/12/17/addressing-barriers-to-responsible-innovation/

Boeschoten, L., Ausloos, J., Moeller, J., Araujo, T., & Oberski, D. L. (2020). *Digital trace data collection through data donation* (arXiv:2011.09851). arXiv. https://doi.org/10.48550/arXiv.2011.09851

Boucher, V. (2022). *Down the TikTok Rabbit Hole: Testing the TikTok Algorithm's Contribution to Right Wing Extremist Radicalization* [Thesis]. https://qspace.library.queensu.ca/handle/1974/30197

Brady, W. J., Gantman, A. P., & Van Bavel, J. J. (2020). Attentional capture helps explain why moral and emotional content go viral. *Journal of Experimental Psychology: General*, *149*, 746–756. https://doi.org/10.1037/xge0000673

Brodersen, K., & Hauser, A. (2017). *CausalImpact: An R package for causal inference using Bayesian structural time-series models*. https://google.github.io/CausalImpact/CausalImpact.html

Brown, M. A., Bisbee, J., Lai, A., Bonneau, R., Nagler, J., & Tucker, J. A. (2022). *Echo Chambers, Rabbit Holes, and Algorithmic Bias: How YouTube Recommends Content to Real Users* (SSRN Scholarly Paper No. 4114905). https://doi.org/10.2139/ssrn.4114905

Bruni, L., Francalanci, C., & Giacomazzi, P. (2012). The Role of Multimedia Content in Determining the Virality of Social Media Information. *Information*, *3*(3), Article 3. https://doi.org/10.3390/info3030278

Burke, M., Cheng, J., & de Gant, B. (2020). Social Comparison and Facebook: Feedback, Positivity, and Opportunities for Comparison. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3313831.3376482

Bustamante, C. M. V. (2022). *Social Media Recommendation Algorithms* (Technology and Public Purpose Project). Belfer Center for Science and International Affairs Harvard Kennedy School.

Cameron, D., Wodinsky, S., DeGeurin, M., & Germain, T. (2022, April 18). *Read the Facebook Papers for Yourself*. Gizmodo. https://gizmodo.com/facebook-papers-how-to-read-1848702919

Carroll, M., Hadfield-Menell, D., Russell, S., & Dragan, A. (2021). Estimating and Penalizing Preference Shift in Recommender Systems. *Fifteenth ACM Conference on Recommender Systems*, 661–667. https://doi.org/10.1145/3460231.3478849

Chowdhury, R., & Gonzales, A. (2022). *Investing in privacy enhancing tech to advance transparency in ML*. https://blog.twitter.com/engineering/en_us/topics/insights/2022/investing-in-privacy-enhancing-tech-to-advance-transparency-in-ML

Christakopoulou, K., & Banerjee, A. (2019). Adversarial attacks on an oblivious recommender. *Proceedings of the 13th ACM Conference on Recommender Systems*, 322–330. https://doi.org/10.1145/3298689.3347031

Community Guidelines Enforcement Report. (2023, March 27). TikTok. https://www.tiktok.com/transparency/en-us/community-guidelines-enforcement-2022-4/

Coons, C., Portman, R., Klobuchar, A., & Cassidy, B. (2022). *Platform Accountability and Transparency Act*. https://www.coons.senate.gov/imo/media/doc/text_pata_117.pdf

Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2022). Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1571–1583. https://doi.org/10.1145/3531146.3533213

Crocker, A., Gebhart, G., Mackay, A., Opsahl, K., Tsukayama, H., Williams, J. L., & York. (2019, June 12). *Who Has Your Back? Censorship Edition 2019*. Electronic Frontier Foundation. https://www.eff.org/wp/who-has-your-back-2019

Csikszentmihalyi, M., & Larson, R. (2014). Validity and Reliability of the Experience-Sampling Method. In M. Csikszentmihalyi (Ed.), *Flow and the Foundations of Positive Psychology: The*

Collected Works of Mihaly Csikszentmihalyi (pp. 35–54). Springer Netherlands. https://doi.org/10.1007/978-94-017-9088-8_3

Cunningham, S. (2021). *Causal Inference: The Mixtape*. Yale University Press. https://mixtape.scunning.com/

Dash, A., Mukherjee, A., & Ghosh, S. (2019). A Network-centric Framework for Auditing Recommendation Systems. *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, 1990–1998. https://doi.org/10.1109/INFOCOM.2019.8737486

Davey, J., Comerford, M., Guhl, J., Baldet, W., & Colliver, C. (2021). *Post-Organisational Violent Extremist & Terrorist Content*.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), Article 1. https://doi.org/10.1609/icwsm.v11i1.14955

Davis, A. (2019, August 1). Open-Sourcing Photo- and Video-Matching Technology to Make the Internet Safer. *Meta*. https://about.fb.com/news/2019/08/open-source-photo-video-matching/

Digital Regulation Cooperation Forum. (2022, September 23). *Auditing algorithms: The existing landscape, role of regulators and future outlook*. GOV.UK. https://www.gov.uk/government/publications/findings-from-the-drcf-algorithmic-processing-workstream-spring-2022/auditing-algorithms-the-existing-landscape-role-of-regulators-and-future-outlook

do Prado, K. S. (2023). *Awesome MLOps* [Python]. https://github.com/kelvins/awesome-mlops

Do, V., Corbett-Davies, S., Atif, J., & Usunier, N. (2022). Online Certification of Preference-Based Fairness for Personalized Recommender Systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(6), Article 6. https://doi.org/10.1609/aaai.v36i6.20606

Dwoskin, E. (2018, August 27). Facebook is rating the trustworthiness of its users on a scale from zero to 1. *Washington Post*. https://www.washingtonpost.com/technology/2018/08/21/facebook-is-rating-trustworthiness-its-users-scale-zero-one/

eSafety Commissioner. (2023). *Position statement: Recommender systems and algorithms*. https://www.esafety.gov.au/industry/tech-trends-and-challenges/recommender-systems-and-algorithms

Etic Lab. (2018, July 16). Six Degrees of Jordan Peterson. *Etic Lab*. https://eticlab.co.uk/jordan-petersons-proliferation/

European Union, Article 37, 277 OJ L (2022). http://data.europa.eu/eli/reg/2022/2065/oj/eng

Evans, C., & Kasirzadeh, A. (2022). *User Tampering in Reinforcement Learning Recommender Systems* (arXiv:2109.04083). arXiv. http://arxiv.org/abs/2109.04083

Facebook Papers. (2019). *The Meaningful Social Interactions Metric Revisited: Part 2*. https://s3.documentcloud.org/documents/21602131/tier0_rank_exp_1119.pdf

Facebook Papers. (2020). *MSI Metric Changes for 2020H1*. https://s3.documentcloud.org/documents/21601827/tier2_rank_ro_0120.pdf

Facebook Papers. (2023). *The Facebook Papers*. The Facebook Papers. https://facebookpapers.com/

Fernandez, M., Bellogin, A., & Cantador, I. (2020). Analyzing the Effect of Recommendation Algorithms on the Amplification of Misinformation. *ArXiv*. https://arxiv.org/pdf/2103.14748v1.pdf

Fischer, S. (2022, August 16). *Scoop: Oracle begins auditing TikTok's algorithms*. Axios. https://www.axios.com/2022/08/16/oracle-auditing-tiktok-algorithms

Fredheim, R., Bay, S., Dek, A., Biteniece, N., Gallacher, J., Bertolin, G., Christie, E., Kononova, C., & Marchenko, T. (2019). *Responding to Cognitive Security Challenges*. https://stratcomcoe.org/publications/responding-to-cognitive-security-challenges/113

Gallacher, J., & Bright, J. (2021). *Hate Contagion: Measuring the spread and trajectory of hate on social media*. PsyArXiv. https://doi.org/10.31234/osf.io/b9qhd

Gallacher, J., & Heerdink, M. (2019). Measuring the effect of Russian Internet Research Agency information operations in online conversations. *Defence Strategic Communications*, *6*(1). http://stratcomcoe.org/publications/measuring-the-effect-of-russian-internet-research-agency-information-operations-in-online-conversations/95

Gausen, A., Luk, W., & Guo, C. (2022). Using Agent-Based Modelling to Evaluate the Impact of Algorithmic Curation on Social Media. *Journal of Data and Information Quality*, *15*(1), 2:1-2:24. https://doi.org/10.1145/3546915

GCHQ. (2023). *Bailo* [TypeScript]. GCHQ. https://github.com/gchq/Bailo

Ge, Y., Liu, S., Fu, Z., Tan, J., Li, Z., Xu, S., Li, Y., Xian, Y., & Zhang, Y. (2022). *A Survey on Trustworthy Recommender Systems* (arXiv:2207.12515). arXiv. http://arxiv.org/abs/2207.12515

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*.

Gillespie, T. (2018). Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. In *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (p. 288). https://doi.org/10.12987/9780300235029

Goel, S., Anderson, A., Hofman, J., & Watts, D. J. (2016). The Structural Virality of Online Diffusion. *Management Science*, *62*(1), 180–196. https://doi.org/10.1287/mnsc.2015.2158

Google. (2018). *Fighting child sexual abuse online*. https://protectingchildren.google/intl/en_uk/

Google. (2023). *Code of Practice on Disinformation – Repo of Google for the period 1 July 2022—30 September 2022*. https://disinfocode.eu/reports-archive/?years=2023

Gordon, M. L., Lam, M. S., Park, J. S., Patel, K., Hancock, J. T., Hashimoto, T., & Bernstein, M. S. (2022). Jury Learning: Integrating Dissenting Voices into Machine Learning Models. *CHI Conference on Human Factors in Computing Systems*, 1–19. https://doi.org/10.1145/3491102.3502004

Gordon, M. L., Zhou, K., Patel, K., Hashimoto, T., & Bernstein, M. S. (2021). The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3411764.3445423

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), 2053951719897945. https://doi.org/10.1177/2053951719897945

Griffioen, N., Van Rooij, M. M. J. W., Lichtwarck-Aschoff, A., & Granic, I. (2020). A Stimulated Recall Method for the Improved Assessment of Quantity and Quality of Social Media Use. *Journal of Medical Internet Research*, *22*(1), e15529. https://doi.org/10.2196/15529

Hansen, C., Mehrotra, R., Hansen, C., Brost, B., Maystre, L., & Lalmas, M. (2021, March 8). *Shifting Consumption towards Diverse content via Reinforcement Learning*. Spotify Research. https://research.atspotify.com/2021/03/shifting-consumption-towards-diverse-content-via-reinforcement-learning/

Harman, M. (2020, July 23). *A Facebook-scale simulator to detect harmful behaviors*. https://ai.facebook.com/blog/a-facebook-scale-simulator-to-detect-harmful-behaviors/

Haroon, M., Chhabra, A., Liu, X., Mohapatra, P., Shafiq, Z., & Wojcieszak, M. (2022). *YouTube, The Great Radicalizer? Auditing and Mitigating Ideological Biases in YouTube Recommendations* (arXiv:2203.10666). arXiv. http://arxiv.org/abs/2203.10666

Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, M. (2007). *Experience Sampling Method: Measuring the Quality of Everyday Life*. SAGE.

Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Martinez-Hernandez, V., Sanchez, V., & Perez-Meana, H. (2018). *A Web Scraping Methodology for Bypassing Twitter API Restrictions* (arXiv:1803.09875). arXiv. http://arxiv.org/abs/1803.09875

Hoang, T.-A., & Lim, E.-P. (2016). Tracking Virality and Susceptibility in Social Media. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 1059–1068. https://doi.org/10.1145/2983323.2983800

Hobbs, T. D., Barry, R., & Koh, Y. (2021, December 17). 'The Corpse Bride Diet': How TikTok Inundates Teens With Eating-Disorder Videos. *Wall Street Journal*. https://www.wsj.com/articles/how-tiktok-inundates-teens-with-eating-disorder-videos-11639754848

Holbrook, D. (2015). Designing and Applying an 'Extremist Media Index'. *Perspectives on Terrorism*, *9*(5), 57–68.

Holbrook, D. (2017). *What Types of Media Do Terrorists Collect?: An Analysis of Religious, Political, and Ideological Publications Found in Terrorism Investigations in the UK*. International Centre for Counter-Terrorism. https://www.jstor.org/stable/resrep29419

Hosseinmardi, H., Ghasemian, A., Clauset, A., Mobius, M., Rothschild, D. M., & Watts, D. J. (2021). Examining the consumption of radical content on YouTube. *Proceedings of the National Academy of Sciences*, *118*(32), e2101967118. https://doi.org/10.1073/pnas.2101967118

Huszár, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjer, A., & Hardt, M. (2022). Algorithmic amplification of politics on Twitter. *Proceedings of the National Academy of Sciences*, *119*(1), e2025334119. https://doi.org/10.1073/pnas.2025334119

Hutton, L., & Henderson, T. (2015). Making Social Media Research Reproducible. *Proceedings of the International AAAI Conference on Web and Social Media*, *9*(4), Article 4. https://doi.org/10.1609/icwsm.v9i4.14685

Hyers, L. L. (2018). *Diary Methods*. Oxford University Press.

Iyer, R. (2023, January 10). When should companies optimize for engagement? [Substack newsletter]. *The Psychology of Technology Institute Newsletter*. https://psychoftech.substack.com/p/when-should-companies-optimize-for

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). Springer. https://doi.org/10.1007/978-1-4614-7138-7

Jay, M. (2023). *PySift* [Python]. OpenMined. https://github.com/OpenMined/PySyft (Original work published 2017)

Jiang, J. A., Scheuerman, M. K., Fiesler, C., & Brubaker, J. R. (2021). Understanding international perceptions of the severity of harmful content online. *PLOS ONE*, *16*(8), e0256762. https://doi.org/10.1371/journal.pone.0256762

Jiang, S., & Ngien, A. (2020). The Effects of Instagram Use, Social Comparison, and Self-Esteem on Social Anxiety: A Survey Study in Singapore. *Social Media + Society*, *6*(2), 2056305120912488. https://doi.org/10.1177/2056305120912488

Jigsaw. (2021). *Perspective API*. https://perspectiveapi.com/

Kennedy, T. (2020). *Indigenous Peoples' Experiences of Harmful Content on Social Media*. https://research-management.mq.edu.au/ws/portalfiles/portal/135775224/MQU_HarmfulContentonSocialMedia_report_201202.pdf

Knijnenburg, B. P., & Willemsen, M. C. (2015). Evaluating Recommender Systems with User Experiments. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender Systems Handbook* (pp. 309–352). Springer US. https://doi.org/10.1007/978-1-4899-7637-6_9

Koh, Y. (2022). *How YouTube Kids Cleaned Up Its Act—WSJ*. https://www.wsj.com/articles/how-youtube-kids-cleaned-up-its-act-11646476200

Kohlbrenner, A., Kaiser, B., Kandula, K., Weiss, R., Mayer, J., Han, T., & Helmer, R. (2022). *Rally and WebScience: A Platform and Toolkit for Browser-Based Research on Technology and Society Problems* (arXiv:2211.02274). arXiv. http://arxiv.org/abs/2211.02274

Krueger, D., Maharaj, T., & Leike, J. (2020). *Hidden Incentives for Auto-Induced Distributional Shift* (arXiv:2009.09153). arXiv. http://arxiv.org/abs/2009.09153

Lazar, S. (Director). (2023, January 26). *Lecture II: Communicative Justice and the Distribution of Attention*. https://www.youtube.com/watch?v=97U8BZAbJYo

Leqi, L., & Dean, S. (2022). *Engineering a Safer Recommender System*. Workshop on Responsible Decision Making in Dynamic Environments, Baltimore, Maryland.

Lorenz, T. (2022, April 11). Internet 'algospeak' is changing our language in real time, from 'nip nops' to 'le dollar bean'. *Washington Post*. https://www.washingtonpost.com/technology/2022/04/08/algospeak-tiktok-le-dollar-bean/

Mac, R. (2021). Engagement ranking boost, M.S.I., and more. *The New York Times*.

Measuring Our Progress Combating Hate Speech. (2020, November 19). *Meta*. https://about.fb.com/news/2020/11/measuring-progress-combating-hate-speech/

Menn, J. (2023). *Russian propgandists get Twitter verification two years in Ukraine war—The Washington Post*. https://www.washingtonpost.com/technology/2023/02/22/russian-propagandists-said-buy-twitter-blue-check-verifications/

Meserole, C. (2022). *How do recommender systems work on digital platforms?* (TechStream). Brookings Institute.

Meßmer, A.-K., & Degeling, M. (2023). *Auditing Recommender Systems: Putting the DSA into practice with a risk-scenario-based approach* (Strengthening the Digital Public Sphere and Platform Regulation). Stiftung Neue Verantwortung. https://www.stiftung-nv.de/de/publication/auditing-recommender-systems

Meta Launches New Content Moderation Tool as It Takes Chair of Counter-Terrorism NGO. (2022, December 13). *Meta*. https://about.fb.com/news/2022/12/meta-launches-new-content-moderation-tool/

Metz, R. (2021). Likes, anger emojis and RSVPs: The math behind Facebook's News Feed—And how it backfired. *CNN Business*.

Microsoft. (2023). *Machine Learning Operations (MLOps)*. https://azure.microsoft.com/en-gb/products/machine-learning/mlops/

Milano, S., Taddeo, M., & Floridi, L. (2020). Recommender systems and their ethical challenges. *AI & SOCIETY*, *35*(4), 957–967. https://doi.org/10.1007/s00146-020-00950-y

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. https://doi.org/10.1145/3287560.3287596

Morley, J., Kinsey, L., Elhalal, A., Zoisi, M., & Floridi, L. (2021). Operationalizing AI ethics: Barriers, enablers, and next steps. *AI & Society*.

NCSC. (2022). *Principles for the security of machine learning*. https://www.ncsc.gov.uk/collection/machine-learning/requirements-and-development/model-architecture

Nishimoto, B. E. (2021, July 20). Multi-armed Bandits: An alternative to A/B testing. *Medium*. https://medium.com/@brunonishimoto/multi-armed-bandits-an-alternative-to-a-b-testing-8acce8e12549

O'Connor, J. (2021, April 6). *Building greater transparency and accountability with the Violative View Rate*. YouTube Official Blog. https://blog.youtube/inside-youtube/building-greater-transparency-and-accountability/

Ofcom. (2022). *The Buffalo Attack: Implications for Online Safety*.

Ovadya, A., & Thorburn, L. (2023). *Bridging Systems: Open Problems for Countering Destructive Divisiveness across Ranking, Recommenders, and Governance* (arXiv:2301.09976). arXiv. http://arxiv.org/abs/2301.09976

Phellas, C. N., Bloch, A., & Seale, C. (2011). STRUCTURED METHODS: INTERVIEWS, QUESTIONNAIRES AND OBSERVATION. *DOING RESEARCH*.

Raji, I. D., Xu, P., Honigsberg, C., & Ho, D. E. (2022). *Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance* (arXiv:2206.04737). arXiv. http://arxiv.org/abs/2206.04737

Rakova, B. (2021). Challenges for Responsible AI Practitioners and the Importance of Solidarity. *Partnership on AI*.

Ramaciotti Morales, P., & Cointet, J.-P. (2021). Auditing the Effect of Social Network Recommendations on Polarization in Geometrical Ideological Spaces. *Fifteenth ACM Conference on Recommender Systems*, 627–632. https://doi.org/10.1145/3460231.3478851

Ribeiro, M. H., Ottoni, R., West, R. Almeida, V. A. and Meira, W. (2019). Auditing Radicalization Pathways on YouTube. https://arxiv.org/pdf/1908.08313.pdf

Ribeiro, M. H., Veselovsky, V., & West, R. (2023). *The Amplification Paradox in Recommender Systems* (arXiv:2302.11225). arXiv. http://arxiv.org/abs/2302.11225

Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to Recommender Systems Handbook. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender Systems Handbook* (pp. 1–35). Springer US. https://doi.org/10.1007/978-0-387-85820-3_1

Rogers, E. (2022). The Role of User Agency in the Algorithmic Amplification of Terrorist and Violent Extremist Content. *GNET*. https://gnet-research.org/2022/09/21/the-role-of-user-agency-in-the-algorithmic-amplification-of-terrorist-and-violent-extremist-content/

Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. B. (2021). HateCheck: Functional Tests for Hate Speech Detection Models. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 41–58. https://doi.org/10.18653/v1/2021.acl-long.4

Saito, Y., & Joachims, T. (2022). Counterfactual Evaluation and Learning for Interactive Systems: Foundations, Implementations, and Recent Advances. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4824–4825. https://doi.org/10.1145/3534678.3542601

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). *Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms*.

Schmon, C., & Pedersen, H. (2022, May 19). *Platform Liability Trends Around the Globe: From Safe Harbors to Increased Responsibility*. Electronic Frontier Foundation. https://www.eff.org/deeplinks/2022/05/platform-liability-trends-around-globe-safe-harbors-increased-responsibility

Schwemer, S. F. (2021). *Recommender Systems in the EU: from Responsibility to Regulation?* FAccTRec Workshop, Amsterdam, Netherlands.

Singh, S., & Doty, L. (2021). *The Transparency Report Tracking Tool: How Internet Platforms Are Reporting on the Enforcement of Their Content Rules*. Open Technology Institute.

Slivkins, A. (2022). *Introduction to Multi-Armed Bandits* (arXiv:1904.07272). arXiv. http://arxiv.org/abs/1904.07272

Srba, I., Moro, R., Tomlein, M., Pecher, B., Simko, J., Stefancova, E., Kompan, M., Hrckova, A., Podrouzek, J., Gavornik, A., & Bielikova, M. (2022). Auditing YouTube's Recommendation Algorithm for Misinformation Filter Bubbles. *ACM Transactions on Recommender Systems*. https://doi.org/10.1145/3568392

Srivastava, A. K., & Mishra, R. (2023). Analyzing Social Media Research: A Data Quality and Research Reproducibility Perspective. *IIM Kozhikode Society & Management Review*, *12*(1), 39–49. https://doi.org/10.1177/22779752211011810

Stiftung Neue Verantwortung. (2022, January 26). *Approaches to Analyse and Evaluate AI-Based Recommendation Systems for Internet Intermediaries*. https://www.stiftung-nv.de/en/subproject/approaches-analyse-and-evaluate-ai-based-recommendation-systems-internet-intermediaries

Stray, J., Halevy, A., Assar, P., Hadfield-Menell, D., Boutilier, C., Ashar, A., Beattie, L., Ekstrand, M., Leibowicz, C., Sehat, C. M., Johansen, S., Kerlin, L., Vickrey, D., Singh, S., Vrijenhoek, S., Zhang, A., Andrus, M., Helberger, N., Proutskova, P., … Vasan, N. (2022). *Building Human Values into Recommender Systems: An Interdisciplinary Synthesis* (arXiv:2207.10192). arXiv. https://doi.org/10.48550/arXiv.2207.10192

Tech Against Terrorism. (2021). *Position paper: Content personalisation and the online dissemination of terrorist and violent extremist content*. https://www.techagainstterrorism.org/wp-

content/uploads/2021/06/210120-TAT-Position-Paper-content-personalisation-and-online-dissemination-of-terrorist-content-JB-vFINAL-DA.pdf?__cf_chl_tk=8xBYhAQzQNTzyPrdw1WIry5Cqs1UVkIP2EFSuIlFodU-1674738919-0-gaNycGzNCD0

Thorburn, L., Stray, J., & Bengani, P. (2022a, May 11). What Does it Mean to Give Someone What They Want? The Nature of Preferences in Recommender Systems. *Understanding Recommenders*. https://medium.com/understanding-recommenders/what-does-it-mean-to-give-someone-what-they-want-the-nature-of-preferences-in-recommender-systems-82b5a1559157

Thorburn, L., Stray, J., & Bengani, P. (2022b, October 11). Is Optimizing for Engagement Changing Us? *Understanding Recommenders*. https://medium.com/understanding-recommenders/is-optimizing-for-engagement-changing-us-9d0ddfb0c65e

Thorburn, L., Stray, J., & Benghani. (2022, November 23). How to Measure the Causal Effects of Recommenders. *Understanding Recommenders*. https://medium.com/understanding-recommenders/how-to-measure-the-causal-effects-of-recommenders-5e89b7363d57

Thorley, T., Llansó, E., & Meserole, C. (2022). Methodologies to Evaluate Content Sharing Algorithms & Processes. *GIFCT Technical Approaches Working Group*.

Thorn. (2020). *Safer: Building the internet we deserve.* Safer: Building the Internet We Deserve. https://safer.io/

Tracking Exposed. (2022). *Shadow-promotion: TikTok's algorithmic recommendation of banned content in Russia*. https://tracking.exposed/pdf/tiktok-russia-ShadowPromotion.pdf

Tracking Exposed. (2023). *Tracking Exposed toolkit* [HTML]. tracking-exposed. https://github.com/tracking-exposed/trex (Original work published 2018)

Trask, A. (2022, January 20). *Announcing our partnership with Twitter to advance algorithmic transparency*. OpenMined Blog. https://blog.openmined.org/announcing-our-partnership-with-twitter-to-advance-algorithmic-transparency/

Twitter Transparency Center. (2021). https://transparency.twitter.com/en.html

Uk Catapult. (2021). *Challenges to Responsible AI Adoption in Industry*. https://www.digicatapult.org.uk/wp-content/uploads/2021/11/Digital_Catapult_Challenges_to_Responsible_AI_Adoption_Report_Juy_2021.pdf

Valkenburg, P. M., Peter, J., & Walther, J. B. (2016). Media Effects: Theory and Research. *Annual Review of Psychology*, *67*(1), 315–338. https://doi.org/10.1146/annurev-psych-122414-033608

van Berkel, N., Ferreira, D., & Kostakos, V. (2017). The Experience Sampling Method on Mobile Devices. *ACM Computing Surveys*, *50*(6), 93:1-93:40. https://doi.org/10.1145/3123988

Vincent, J. (2020, July 23). *Facebook is simulating users' bad behavior using AI*. The Verge. https://www.theverge.com/2020/7/23/21333854/facebook-ai-simulation-bad-behavior-ww-web-base-simulator

Waseem, Z. (2016). Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. *Proceedings of the First Workshop on NLP and Computational Social Science*, 138–142. https://doi.org/10.18653/v1/W16-5618

Waseem, Z., & Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. *Proceedings of the NAACL Student Research Workshop*, 88–93. https://doi.org/10.18653/v1/N16-2013

Waters, G. & Postings, R. (2018). Spiders of the caliphate: Mapping the Islamic State's global support network on Facebook. Counter Extremism Project. https://www.counterextremism.com/press/icymi-cep-report-reveals-how-isis-supporters-are-organizing-facebook

Whittaker, J. (2022). *Recommendation Algorithms and Extremist Content: A Review of Empirical Evidence* (GIFCT WORKING GROUPS OUTPUT 2022, pp. 162–191) [GIFCT Transparency Working Group]. GIFCT. https://gifct.org/wp-content/uploads/2022/09/GIFCT-22WG-Combined-US-Sizing2.1-2.pdf

Whittaker, J., Looney, S., Reed, A., & Votta, F. (2021). Recommender systems and the amplification of extremist content. *Internet Policy Review*, *10*(2). https://doi.org/10.14763/2021.2.1565

Young, S. W. H. (2014). Improving Library User Experience with A/B Testing: Principles and Process. *Weave: Journal of Library User Experience*, *1*(1). https://doi.org/10.3998/weave.12535642.0001.101

Zhang, S. (2021). *Measuring Algorithmic Bias in Job Recommender Systems: An Audit Study Approach*.

Zhang, Y., Xiao, J., Hao, S., Wang, H., Zhu, S., & Jajodia, S. (2020). Understanding the Manipulation on Recommender Systems through Web Injection. *IEEE Transactions on Information Forensics and Security*, *15*, 3807–3818. https://doi.org/10.1109/TIFS.2019.2954737

Zhou, X., & Li, Y. (2021). Large-Scale Modeling of Mobile User Click Behaviors Using Deep Learning. *Fifteenth ACM Conference on Recommender Systems*, 473–483. https://doi.org/10.1145/3460231.3474264

9 Appendices

Analysis Method

The first stage of our analysis involved working closely with the Ofcom team to understand the exact scope of the deliverables and milestones. We organised multiple workshops at the outset of the project to ensure we would align closely with this. Concurrently, we began booking experts from academia, government and industry to secure the interviews needed for our accelerated timeline.

The second stage involved gathering and analysing evidence through a multi-method research design. Our team conducted extensive research into cutting-edge social, computer and policy science literature. We were able to access to the latest scholarly preprints, advance publications and working papers from academic conference proceedings where the most recent research findings appear. Additionally, we collected evidence from extensive in-depth interviews and stakeholder roundtables, employing semi-structured interviews. Once we reached saturation in the collection of evidence from both literature and interviews, we coded the evidence into themes and catalogued the evidence in association with our guiding research questions. Using iterative coding, this methodology revealed consistent findings and salient trends, and identified blind spots in both policy and research.

Finally, we submitted the entire assessment for a rapid review by independent experts and incorporated several waves of feedback from the Ofcom team. This ensured the highest standards in research methods, captured the latest evidence from policy-relevant and applied research, and verified and validated our findings and analysis.

Multi-method research design such as this provides the most comprehensive form of applied policy analysis. The most important methods are likely to be qualitative (expert consultations with experienced staff in civil society, industry and government), comparative (across different types of systems, and across the most widely used platforms) and quantitative (systematic reviews, meta-analysis of data).

While the use of opinion data on how the public understands recommender systems and thinks about the regulatory environment is not specified in the tender request, we believe that going forward it is important that policymakers have full sight of the public appetite for regulatory solutions. In this way, we are able to provide the most thorough review of the latest research, with multiple forms of evidence, drawn from several countries and platforms, and with the added value of real measures of public understanding of recommender systems.

One risk to this research is in the use of interview methods in the contemporary work environment. For our expert consultations we found that health crises, time zones and changing work commitments meant that flexible interviewing arrangements, including some email-based follow-up questions, elicited higher response rates than requests for in-person interviews. The timeline was relatively tight, so to mitigate this we planned for virtual expert interviews, supplemented with email surveys, in our expert consultations. The second risk we faced was that the platforms examined may have adapted their recommender systems while we were conducting our research. Our mitigation strategy was to include in our literature review the latest research we were able to access as academics (working papers and conference proceedings), to have multiple forms of qualitative, comparative and quantitative data, and to have a commissioned, independent review at the final stages of preparing deliverables for Ofcom. It is possible, indeed likely, that recommender systems

have evolved since the completion of the report, but we remain confident that the broad trends and general findings we identify will remain accurate and applicable for some time to come.

We ensured GDPR compliance when collecting data, including through interviews, and used encrypted communication channels and data storage.

B - List of expert interviews and workshops

Table 2 Expert participants across interviews and workshops.

Pseudonym	Type	Affiliation type(s)	Date	No. Participants
A6	Interview	CSO, Academia	22 Nov 2022	1
A1	Interview	Academia	28 Nov 2022	1
Ind1	Interview	CSO, Industry	28 Nov 2022	1
A2	Interview	Academia	30 Nov 2022	1
CS1	Interview	CSO	1 Dec 2022	1
A3	Interview	Academia	2 Dec 2022	1
CS2	Interview	CSO, Academia	5 Dec 2022	1
CS3	Interview	Academia, CSO, Industry	6 Dec 2022	1
CS4	Interview	CSO	7 Dec 2022	1
Ind2	Interview	Academia, Industry	8 Dec 2022	1
A4	Interview	CSO, Industry	8 Dec 2022	1
Ind3	Interview	Industry, CSO	8 Dec 2022	1
DW1	Workshop	Industry	9 Dec 2022	3
DW2	Workshop	CSO, Industry, Academia	12 Dec 2022	3
DW3	Workshop	Industry	13 Dec 2022	2
Ind4	Interview	Academia, Industry	13 Dec 2022	1
CS5	Interview	CSO, Academia	14 Dec 2022	1
CS6	Interview	CSO	15 Dec 2022	1
A5	Interview	Academia, CSO	16 Dec 2022	1
Gov1	Interview	Government	22 Dec 2022	2
A7	Interview	Industry, Academia	19 Dec 2022	1
Ind5	Interview	Industry	27 Jan 2023	4

C Data science discovery workshop case study

Introduction to the case study

- Participants are to imagine that they are now data scientists working for a medium-sized social media platform “Flapper”. Flapper lets users share text, images, and videos with each other via “flaps”. The network is set up such that users can see the content that users in their immediate network are sharing, as well as users in their wider network (friends of friends). In addition to this, some content from the wider network (un-connected) can be discovered and viewed.
- Due to an increase in the amount of content shared, Flapper has recently started to use a range of recommender systems to rank/promote the content which is displayed to users, as well as suggest users to follow (and auto-complete search terms). This is to try and ensure that the most relevant and interesting content is shown to users first.
- Flapper has received notice that these recommender systems might be playing a role in promoting illegal content to users.
- Without going too far into the definitions, we can think of illegal content as broadly covering terrorist/violent extremist content or child abuse material, etc.

Workshop discussion topic

As data scientists at the platform, and with access to the data and resources that a platform such as this is likely to have, our question is how Flapper should assess whether their recommender systems are indeed promoting illegal content to users:

- How can we assess where illegal content is ranked relative to legal content?
- How can this illegal content be detected?
- How can we assess whether their systems are affecting the “virality” of illegal content?
- How can we measure speed and network spread
- How can they assess whether their systems are creating pathways from harmful to illegal content?

D Comparative methods table

Table 3 Methods scoring across methodological elements.

Reference	Method	Insight	Resources	Cost	Validity	Standardisation
B1	Prevalence	4.0	4.0	4.0	5.0	8.0
B2	Virality	5.5	5.0	5.5	5.0	3.0
B3	Pathways	7.5	5.0	5.0	5.0	2.5
B4	Real-world outcomes	10	9.0	9.0	5.0	5.0
C1	Speculative	2.5	4.0	4.0	2.5	3.0
C2	Descriptive	5.5	4.0	3.5	8.5	8.0
C3	Causal	10	8.0	8.0	8.5	5.0
D1	Machine learning classifiers	8.0	7.0	3.5	6.5	9.0
D2	User reports	5.0	3.0	1.0	3.0	8.5
D3	User surveys	6.0	4.0	3.0	5.0	8.0
D4	Civil society reports	4.0	2.0	2.5	2.5	1.0
E1	On-platform experiment	9.0	4.0	8.5	9.0	2.0
E2	Off-platform experiment	5.0	8.0	9.5	6.5	2.0
E3	Causal inference on observational data	8.0	7.5	3.0	7.5	2.0
E4	Recommender system debugging	9.5	9.0	5.0	8.5	3.0
F1	Survey instruments	5.0	4.3	3.5	4.3	4.0
F2	Experience sampling	7.0	4.0	4.0	7.0	4.0
F3	Diary studies	6.0	6.0	6.0	4.0	2.0
F4	Stimulated recall	6.5	6.0	7.0	5.0	2.0
G1	Whole platform simulations	4.0	9.0	9.0	3.0	1.0
G2	Sock-puppet accounts	5.0	6.0	4.0	3.5	2.5
G3	Functional testing	4.0	8.0	2.0	5.0	3.5
H1	Number of shares or reposts	4.0	4.0	3.0	5.0	6.0
H2	Structural Validity	5.0	5.0	4.0	6.0	5.0
H3	Probability of being shared	6.5	5.0	4.5	4.0	2.0
H4	Epidemiological model	9.0	8.0	5.0	6.0	2.0
I1	Recommendation maps	5.0	5.0	3.0	2.5	4.0
I2	Distance	4.0	4.0	2.5	2.5	6.5
I3	Learning of unwanted preferences	8.0	6.5	6.5	7.0	3.0
I4	Increasingly extreme content	7.0	6.5	6.5	7.0	4.0
I5	Explore/exploit trade-off	3.0	1.0	1.0	3.0	7.5

Pattrn Analytics & Intelligence (Pattrn.AI) is a spinout company from Oxford University that empowers public agencies and enterprises to effectively detect, understand, and mitigate hostile information campaigns and online harms. Pattrn.AI provides comprehensive insights into global information operations, data flows, and public perception of critical issues. We offer research and consulting services aimed at empowering decision-makers in policy and industry. Our approach combines social data science, machine learning, and social science methodologies to deliver robust and actionable insights. Pattrn.AI was founded by three researchers from the University of Oxford: Philip Howard, Jonathan Bright, and Lisa-Maria Neudert.