Digital Platform Regulators Forum

Literature summary: Harms and risks of algorithms

Working paper

June 2023

Important notice
This document has been prepared by the ACCC, ACMA, eSafety Commissioner and OAIC in their capacity as members of the Digital Platform Regulators Forum (**member regulator**). The information in this publication is for general guidance only. It does not constitute legal or other professional advice, and should not be relied on as a statement of the law in any jurisdiction. Because it is intended only as a general guide, it may contain generalisations. You should obtain professional advice if you have any specific concern.

The member regulators have made every reasonable effort to provide current and accurate information, but do not make any guarantees regarding the accuracy, currency or completeness of that information.

Parties who wish to re-publish or otherwise use the information in this publication must check this information for currency and accuracy prior to publication. This should be done prior to each publication edition, as member regulator guidance and relevant transitional legislation frequently change. Any queries parties have should be addressed to the DP-Reg@oaic.gov.au

Contents

Background

This literature summary is the first output of Digital Platform Regulators Forum (DP-REG)'s joint work in exploring relevant digital platform technologies and their regulatory implications. Each DP-REG member is also separately considering more specific harms stemming from AI relevant to their respective mandates, outlined below.

Background to DP-REG

The Australian Government is taking a wide range of regulatory interventions to help protect Australians online.

To support a streamlined and cohesive approach to regulating digital platforms, the Australian Communications and Media Authority (ACMA), the Australian Competition and Consumer Commission (ACCC), the Office of the Australian Information Commissioner (OAIC), and the Office of the eSafety Commissioner joined forces in March 2022 to establish the Digital Platform Regulators Forum (DP-REG).

DP-REG allows these independent regulators to share information about, and collaborate on, cross-cutting issues and activities on the regulation of digital platforms. This includes consideration of how competition, consumer protection, privacy, online safety and data issues intersect. Where appropriate, DP-REG engages with stakeholders collectively on issues of mutual interest or concern.

DP-REG is not a decision-making body and has no bearing on members' existing regulatory powers, legislative functions or responsibilities. Collaboration under DP-REG is intended to be flexible and recognise the limits of each member's respective regulatory framework. Members are still free to engage bilaterally or outside of DP-REG on issues related to digital platforms.

The current DP-REG governance structure, as outlined in our terms of reference, enables effective cooperation among our regulators at different levels. The heads of each member regulator determine DP-REG's strategic direction, including agreement on the group's annual priorities.

DP-REG's strategic priorities for 2022-23 are outlined in our June 2022 communique. This includes a focus on assessing the impact of algorithms, seeking to improve transparency of digital platforms' activities and how they are protecting users from potential harm, and increased collaboration and capacity building between the four members.

Relevance of algorithms to the remit of DP-REG members

ACMA

The Australian Communications and Media Authority (ACMA) is the independent authority responsible for regulating media and communications in Australia. Most of the entities it regulates use algorithms to deliver content and advertising to Australians, which brings corresponding benefits, risks and challenges. Algorithms are crucial in moderating content and recommending news items. They help broadcasters and streaming services provide targeted advertising and content to users.

However, algorithms can spread and amplify harmful content such as misinformation and disinformation. To address these concerns, the Australian Government plans to grant new powers to ACMA in this area, following the ACMA's oversight of the development and

operation of the Australian Code of Practice on Disinformation and Misinformation since 2020.

Beyond mis- and disinformation, the ACMA takes measures to respond to other sector-specific challenges involving algorithms. This includes monitoring technological solutions to reduce the severity of scams, engaging with stakeholders to understand the role of algorithms in targeted advertising, and conducting consumer research to gain insights into the changing communications and media environment.

ACCC

The Australian Competition and Consumer Commission (ACCC) is an independent Commonwealth statutory agency that promotes competition, fair trading and product safety for the benefit of consumers, businesses and the Australian community. The primary responsibilities of the ACCC are to enforce compliance with the competition, consumer protection, fair trading and product safety provisions of the Competition and Consumer Act 2010 (CCA), regulate national infrastructure, and undertake market inquiries and studies.

The ACCC has been closely considering the competition and consumer impacts of digital platform services over recent years. This includes publishing reports such as the 2019 Final Report in the Digital Platforms Inquiry, the 2021 Final Report in the Digital Advertising Services Inquiry, and the current Digital Platform Services Inquiry, which commenced in February 2020 and is producing six-monthly reports until 2025.

The ACCC recognises algorithms are used in a variety of contexts for many different purposes, bringing both benefits as well as potential risks for competition and consumers. In addition to the relevance of algorithms to the ACCC's current Digital Platform Services Inquiry, the operation of algorithms may be subject to similar types of competition and consumer law considerations that arise in other sectors.

The ACCC has also taken several enforcement actions in cases involving algorithm-related misconduct (for example, see the Trivago case and the iSelect case).

Given the focus of this literature summary on particular types of algorithms most relevant to all DP-REG member regulators, some potential issues posed by algorithms relevant to our remit are not explored in this document (e.g. algorithmic collusion).

OAIC

The Office of the Australian Information Commissioner (OAIC) regulates the Privacy Act, which applies to Australian Government agencies and some private sector organisations. The Act contains 13 Australian Privacy Principles (APPs), which apply across the personal information lifecycle, from collection, through to use and disclosure, storage and destruction.

The APPs are technology-neutral and are designed to adapt to changing and emerging technologies. For example, the obligations in the Privacy Act will apply where personal information is used to train, test or deploy algorithms. This includes obligations to notify individuals about the handling of their personal information, limitations on collecting personal information (including for collection through creation), limitations on use and disclosure of personal information, and providing mechanisms for individuals to access and correct their personal information, among other obligations.

The OAIC publishes guidance that can help entities developing and deploying algorithms that use personal information to identify and mitigate privacy risks and impacts. The Guide to data analytics and the Australian Privacy Principles provides guidance in the data analytics context, while the Guide to undertaking privacy impact assessments provides general guidance on identifying and mitigating privacy risks. In addition, the Australian Privacy Principles guidelines outline the mandatory requirements of the APPs and how the OAIC interprets them.

eSafety Commissioner

eSafety is Australia's national online safety educator and regulator. Its functions include coordinating activities of Commonwealth Departments, authorities and agencies relating to online safety for Australians. eSafety's approach to algorithms is multi-faceted and involves prevention, protection, and proactive and systemic change.

Prevention

eSafety supports and conducts education and community awareness. Recognising the importance of enhancing digital and algorithmic literacy and giving people the skills and confidence to manage their online experiences, eSafety is developing education and training programs to raise awareness of the potential harms associated with recommender systems and the tools to manage them.

Protection

eSafety administers reporting schemes to investigate and act against illegal and restricted online content, non-consensual sharing of intimate images (image-based abuse), cyberbullying material targeting a child, and seriously harmful abuse material targeting an adult. If harmful content goes undetected by algorithm-based moderation systems or if it spreads through recommender systems, individuals can report it to eSafety.

Proactive and systemic change

The *Online Safety Act 2021* enables eSafety to require online service providers to report on how they are meeting the Australian Government's Basic Online Safety Expectations. This includes asking services about their content moderation and recommendation algorithms to improve transparency, accountability, and safety practices.

In June 2023, eSafety introduced enforceable industry codes that require five sections of the online industry to take steps to reduce the availability of seriously harmful online content such as child sexual abuse and pro-terror material. This includes proactive detection requirements on certain social media services. The code obligations come into effect in December 2023.

The industry codes scheme under the Act is a co-regulatory scheme and eSafety is moving to develop industry standards for two sectors of the online industry. The use of systems, technology and/or processes (including algorithms) to detect harmful content plays a crucial role in determining appropriate measures for these standards.

eSafety conducts consultation and horizon scanning to remain focused on the future and ready for emerging issues. This informs the development of position statements on tech trends and challenges. You can find eSafety's position on recommender systems and their algorithms here.

eSafety also supports the online industry to enhance its online safety measures – including content moderation and recommendation algorithms – through its Safety by Design initiative.

Purpose of this literature summary

This literature summary expands and consolidates DP-REG members' understanding of the types of algorithms relevant to their work, and also supports DP-REG's strategic priorities for 2022/23. Desktop research was conducted using resources available to DP-REG member regulators. Deepening our knowledge of these risks can support the future work of individual regulators and of DP-REG.

DP-REG does not claim this paper covers every potential impact of digital platform algorithms or includes every relevant source.

1 Introduction

An algorithm can be described as 'any well-defined computational procedure that takes some value, or set of values, as input and produces some value, or set of values, as output'.[1] Given this broad definition, the use of algorithms is ubiquitous — they determine the contents of our social media feeds, help us find a route home, book accommodation, and support healthcare, education, finance, and beyond.

While algorithms have long been used to aid decision-making online and offline, recent developments such as widespread adoption of machine learning and an exponential increase in available data have created ever-more sophisticated and complex algorithmic decisions that impact our lives more deeply than ever before.[2]

The Digital Platform Regulators Forum (DP-REG) has prepared this literature summary to outline its understanding of the harms and risks posed by some commonly used types of algorithms as at June 2023. Concentrating on consumer/citizen use of digital platform services, the summary focuses on three topics relevant to intersecting priorities of DP-REG members: algorithms used in content moderation, recommender systems and targeted advertising.

Much of the existing research on this topic is specific to certain geographic regions and platforms, which is a key limitation of this summary. Noting the highly US-centric nature of research in this field, DP-REG members have sought to consider literature from other localities to offer a more comprehensive understanding of how algorithms can disproportionately affect non-Western groups.

2 Harms and risks posed by particular categories of algorithms

This section outlines the harms and risks posed by algorithms used in content moderation, recommender systems and targeted advertising. In each of these sections, harms and risks are categorised by whether they affect individuals, society, or both. Further, each harm or risk is categorised under the themes of fairness, trust or safety.[3] In adopting this structure, we note these broad themes are not mutually exclusive and that some harms or risks could be included under several themes.

2.A Harms and risks posed by algorithms used in content moderation

Overview of content moderation systems

To examine algorithmic content moderation effectively, it is important to understand the role algorithms play within wider content moderation ecosystems. Online content moderation is a complex system that includes various components, such as algorithms, platform terms of service, design architectures, human moderators, users, and much more.

Algorithmic content moderation typically involves matching or prediction models contributing to decisions about specific content or accounts. These decisions can include actions such as suspension or removal from a service. However, it can also include informal measures such as 'shadowbanning' or 'downgrading' borderline content (which reduces the exposure of content so fewer people see it[4,5]), deprioritising content and accounts, or excluding them from recommendations made to users.[6]

Given the speed and scale at which material is exchanged online, and expectations for instant sharing, algorithmic moderation has become necessary to create safe and productive spaces for users to engage. Without it, digital platform services risk becoming dysfunctional and toxic, resulting in negative impacts on users and broader communities. The benefits of algorithmic moderation include:

- **Users and communities:** Reducing harmful content and activity in a timely manner and removing illegal content;
- **Digital platform services:** Creating healthier online spaces which can enhance user experiences, trust, inclusivity and ultimately, user retention; and
- **Human moderators:** Minimising harmful content that requires review.

However, a core issue with some algorithmic moderation is the rate of error, commonly referred to as false positives and false negatives. Errors are as much a feature of content moderation as they are a limitation. Some of these technologies rely on probabilistic methods. While these may improve accuracy over non-probabilistic methods, errors are unavoidable.[7] Therefore, there is a trade-off between false positives (mistakenly removing harmless content) and false negatives (failing to remove harmful or misleading information).

These errors can lead to various harms and inequities, including:

- **False positives:** Removal of non-harmful content, such as mislabelling educational content about breastfeeding as pornography; and
- **False negatives:** Disseminating harmful content, such as disinformation presented as credible news.

Errors occur due to a variety of reasons, such as human error, artificial intelligence (AI) misjudgement of context and language, and the limitations of some tools' 'robustness'. Robustness refers to an algorithm's ability to manage circumvention efforts or unexpected inputs that occur when a tool is used in the real world.[8] In some cases, AI may struggle to apply common sense and interpret cultural, linguistic, social and other contexts, which can lead to inaccurate performance. Predictive models are at the greatest risk of missing context because their training data quality and quantity can vary significantly.[9]

However, some types of algorithmic moderation — such as 'hash matching', which identifies content based on unique digital signatures or 'hashes' — have extremely low error rates[10] and can be considered to pose nil or negligible risk when using expert organisations' databases.

2.A.1 Individual impacts

Algorithms have proven to be beneficial in many technologies. However, research and trends data show they also introduce risks to users' online safety. This is particularly evident in social media and other digital platform services where user-generated content is shared.

Content moderation decisions are guided by policies that consider various rights. For example, removing some types of content can impact freedom of expression, while not removing content could harm a person or a group targeted by that content. Every element of a content moderation system can impact rights in complex ways, from deciding whether to keep or remove content to determining what gets amplified or downplayed. These impacts arise from biases, contextual misunderstanding, and the fundamental design choices made by the platform.

Although most major digital platforms use some form of human-based content moderation, there is an increasing reliance on automated processes. This is due to user demands, regulatory pressures, the sheer volume of harmful content circulating online, and the costs of human moderation. While algorithms are a vital tool that allow digital platforms to classify, filter, flag and curate online information at speed, they also have many limitations.[11]

2.A.1.1 Safety: Online harms and impact on users experiencing vulnerability

The creation, sharing and exposure of harmful content online can have serious consequences, causing emotional distress, psychological harm and even physical impacts. This can affect mental health and wellbeing, physical health, personal relationships, self-esteem and confidence, personal safety and financial security — online and offline.[12,13]

Where content moderation systems fail to detect and remove seriously harmful content, it has negative implications in the following areas:

- **Personal safety:** For example, child sexual abuse material that is generated, shared or made accessible, resulting in significant trauma for the victim;
- **Health or wellbeing:** Examples include direct or indirect threats of violence, intimidation, harassment, and promoting graphic violence, eating disorders, and content inappropriate for certain age groups;
- **Personal dignity:** This includes non-consensual sharing of intimate images, sexual extortion, trolling, cumulative or volumetric attacks, bullying, abuse, insults, rumours, and social exclusion;
- **Privacy:** This includes actions such as doxing (publicly releasing confidential information such as an individual's real name, physical address or contact details without consent), sharing intimate images and deepfakes;
- **Participation and free expression:** These rights can be hindered by misinformation or targeted harassment based on hate, racism or other forms that discourage people from engaging online; and
- **Online discrimination:** This involves identity attacks, dehumanisation, hate speech, and sexual harassment/aggression.

Similarly, failing to remove scams from digital platform services can cause harm and pose particular risks to certain groups, including users experiencing vulnerability. In 2021, Indigenous Australians, older Australians, people from culturally and linguistically diverse communities, and people with disability reported record losses to scams.[14]

The ACCC has taken legal action against Facebook owner Meta Platforms, Inc. and Meta Platforms Ireland Limited, alleging that they engaged in false, misleading or deceptive conduct by publishing scam advertisements featuring prominent Australian public figures. It is alleged Meta was aware the celebrity endorsement cryptocurrency scam ads were being displayed on Facebook but did not take sufficient steps to address the issue.[15]

When platforms fail to detect, remove, and hold accounts responsible for violating content, perpetrators are likely to continue or escalate their behaviour.[16]

For related content, see also section **2.B.2.1** (recommender systems) and **2.C.2.2** (targeted advertising).

2.A.1.2 Safety: Exposure to harmful content

Automated content moderation carries a risk: algorithms can unintentionally amplify harmful and extreme content, causing greater harm to people who come across it. If harmful material is not detected and removed, or treated differently, through content moderation processes, it remains in the recommendation pool and may be amplified to others. See also section **2.B.2.2** (recommender systems).

Similarly, exposure to or intentional access of seriously harmful content, such as child sexual abuse imagery and pro-terror material can cause substantial harm. It can affect a person's behaviour and mental health and risks normalising extremely harmful content.[17] Research by the eSafety Commissioner shows 71% of children aged 14-17 have seen sexual images on the internet. This can harm their emotional and mental health.[18]

Algorithms can also expose children to restricted content (X18+ or R18+)[19], such as online pornography, which is not suitable for their age.[20]

2.A.2 Societal impacts

Besides the impacts on the rights and safety of individual users, we must also consider the broader risks to society posed by algorithmic content moderation.

Relying on algorithms to assist in platform content moderation decisions at scale can build or undermine public trust in news and information. Using algorithms in content moderation has the potential to impact broader social structures and democratic institutions, as discussed below.

2.A.2.1 Fairness: Participation in society and promotion of democratic values

In the 1990s, people saw the internet as a great leveller. It would provide universal access to vast quantities of information, offer tools for individuals to produce and distribute their own content, and reduce traditional barriers to entry. It was thought an open and decentralised internet would bypass traditional power structures and lead to a fairer, freer and more equitable society.

While this utopian vision has not fully materialised, the emergence and rapid uptake of social media platforms in the 2000s and 2010s did connect people globally like never before. Of the 5.16 billion global internet users as of January 2023, 4.76 billion use social media services.[21] These new online spaces continue to play a positive role in democratic societies by "increasing the opportunities to exercise individuals' rights such as freedom of expression."[22]

However, digital platforms can maintain significant influence over the nature, circulation and composition of content on their services due to their control — and sometimes inconsistent enforcement — of moderation polices. This has led some academics to refer to them as the "New Governors" of online speech.[23] These issues are amplified by a platform's private interests to retain users and increase engagement to make money.[24]

By relying on a system of increasingly complex and opaque algorithms, platforms can more easily process a high number of "difficult platform governance puzzles"[25] and help reduce the amplification of content that could lead to broader societal harms,[26] such as hate speech and misinformation (which, for the purposes of this paper, includes disinformation). There are obvious concerns about the lack of nuance and the potential for algorithms to stifle genuine discussion and debate — particularly on controversial but important matters. There

is also a potential risk regarding the political and cultural impact that widespread adoption of content moderation algorithms can have in countries outside the United States.

Many of the largest and most popular platforms are headquartered in the United States and have arguably adopted a US-centric worldview. It has been noted that "American lawyers trained and acculturated in American free speech norms and First Amendment Law oversaw the development of content moderation policy".[27] Section 230 of the US Communications Decency Act 1996 (CDA) limits the liability of services for certain types of user-generated content. Recently, there has been debate about how this law applies, especially where platforms moderate or fail to moderate content that has been promoted or uploaded through content filtering.

Caplan notes that even "when [platforms] grow to be the size of Facebook or YouTube, maintaining consistency in decision-making is often at the expense of being localised or contextual… platforms of this size tend to collapse context in favour of establishing global rules that make little sense when applied to content across vastly different cultural and political contexts around the world."[28]

Global platform moderation rules often prioritise freedom of expression, which can be seen as a positive force for protecting individual rights — particularly in countries with stricter speech regulations. However, this emphasis on free speech coupled with an increasing reliance on automated decision making can also result in platforms failing to identify or act against content that can cause serious societal harms. For example, the UN Independent International Fact-Finding Mission on Myanmar found Facebook had played a "determining role" in spreading racial vilification in Myanmar. Facebook recognised that a lack of local knowledge in its US-based moderation and policy teams led to a decision to ban the accounts of an insurgent organisation resisting state-led ethnic cleansing of the Rohingya Muslim minority in the country, as well to prohibit posts that demonstrated support for the organisation.[29] A digital platform adopting a 'one size fits all' approach to how it moderates content globally may fail to recognise the cultural and legal differences that arise between countries.[30] For example, Thailand bans speech that defames or insults the Monarch[31] and Germany specifically prohibits antisemitic hate speech and Holocaust denial.[32]

Platforms can play a meaningful role in promoting democratic culture and values that support freedom of expression. By virtue of their position operating private transnational infrastructure that facilitates online speech, platforms can empower their users and "temper the power of the state to censor".[33] As businesses that succeed based on user and advertiser-loyalty, platforms are also vulnerable to public collective action.[34] However, much of the literature presents a pessimistic view on how platforms use algorithms to moderate content. There is a lack of transparency about the interests and decisions behind algorithms. This lack of transparency raises questions about how these algorithms influence what speech is deemed acceptable or unacceptable. Our analysis of the literature suggests it is important for governments, regulators, platforms and consumers to be aware of the potential power of algorithmic content moderation in shaping public discourse in their countries. They also need to be aware of the threat it can pose to democratic systems and culture.

2.A.2.2 Fairness: Bias and discrimination

Online harms are complex, are often intersectional, and depend on the individual user's unique circumstances. If not used effectively, algorithmic content moderation can further marginalise certain communities and diminish the availability of diverse viewpoints on complex issues. The literature indicates algorithms are trained on existing content that

exhibits human biases such as racial or gender bias. This means the algorithms can perpetuate this bias.

There are documented instances of marginalised communities facing disproportionate reduction and removal of their content, including:[35]

- LGBTQIA+ content creators being 'de-monetised' (losing their ability to earn revenue on a platform) for using the word 'gay';
- TikTok downranking educational content about women's health, pregnancy and menstrual cycles; and
- Marginalised communities needing to use coded language ('algospeak') to avoid detection, such as LGBTQIA+ communities referring to LGBTQIA+ as 'leg booty'.

Predictive content moderation systems enabled by machine learning and neural networks depend on vast amounts of data to make predictions. When the quality and size of data is limited, existing biases can be exacerbated. Visual and audio pose a greater risk,[36] due to issues of data size and quality.[37] Training data may omit examples from certain categories, diverse communities, and languages, leading to erroneous classifications that affect marginalised communities disproportionately.[38] Similarly, humans may introduce their own inherent biases when labelling training data.[39]

Natural language processing (NLP) tools are "typically usually used to parse text in English [and] when parsing non-English text can result in harmful outcomes for non-English speakers, especially when applied to languages that are not very prominent on the internet."[40] Commercial interests driving algorithmic content moderation have resulted in some platforms "effectively exclud[ing] sex workers and marginalis[ing] women and LGBTQIA+ people by removing or restricting their communications."[41] It has also been observed that the 'error' Meta allows in its probabilistic method of content moderation "is not an abstract notion, it is a concrete, lived experience, usually endured by specific (marginalised) groups… an error may represent not only a breach of internal platform law, but also an abrogation of an individual's rights under international human rights law."[42]

Academic studies published in 2019 found that algorithms trained to identify hate speech for removal were more likely to flag social media posts by African Americans discussing contentious events and personal experiences related to racism in the United States.[43] On certain machine learning toxic speech classifiers, observers found over- and under-zealous toxicity predictions. For example, the single-term comment 'Arabs' was classified as 63% toxic, while the phrase 'I love fuhrer' was 3% toxic."[44] Research by the Australian Strategic Policy Institute (ASPI) also suggests TikTok's algorithms apply a heavy-handed approach to global content moderation, seeking to avoid any controversial subjects and topics that may be politically sensitive. For example, TikTok supressed hashtags related to LGBTQIA+ issues in at least 8 languages and used 'shadowbans'.[45] The research also raised concerns about algorithmic decision making by the Chinese government.[46]

Despite its limitations, algorithmic content moderation is an important tool for platforms. This is because human moderators "are doing psychologically scarring work, in sometimes intolerable conditions, often under precarious labour arrangements."[47] However, if not used effectively, algorithmic content moderation can further perpetuate the marginalisation of certain communities and negate a diversity of viewpoints and nuance in complex issues.

Similar concerns may also arise in recommender systems (see **2.B.1.1**) and targeted advertising (see **2.C.1.1**)

2.A.2.3 Trust: Echo chambers, filter bubbles and polarisation

Digital platforms have been part of the news ecosystem for a number of years, hosting news content from traditional media outlets, online publications, and users. Most platform users consume news on these services,[48] and many use them as forums for debate.

However, due to their commercial interests in growth, market dominance and profit, platforms may discourage content that goes against user norms and prioritise filtering only content that meets a user's tastes.[49] As a result, "users will not only be exposed to less diverse content but they will also be less able to post antinormative content as external and internal content moderation policies standardise across platforms."[50] These effects can lead to echo chambers,[51] filter bubbles[52] and polarisation.

The emergence of echo chambers and polarisation in the news environment has various consequences. These include limited exposure to different viewpoints, a greater willingness to spread misinformation, people seeking out information that confirms existing beliefs, and reinforcing existing divisions between ideological-diverse groups. There are fears that limiting individuals' access to contrary viewpoints and reinforcing existing views in this way may cause individual preferences or beliefs to harden over time, and that this may lead to polarisation at a societal level.[53]

Research on these effects is ongoing,[54] but yields mixed results. While "politically partisan online news echo chambers are generally small,"[55] "there is no single uniform trend towards greater polarisation."[56] Some countries have seen a decline in ideological polarisation[57] but an increase in affective polarisation,[58] and News audience polarisation[59] varies greatly."[60] Affective polarisation has contributed to the rise of negative voting (voting against rather than for a party or candidate) in elections, which occurs as a product of the increasing dislike people have for those from the 'other' political party.[61] There are also concerns that echo chambers and filter bubbles could lead to radicalisation or limit the ability of the news media to identify and scrutinise targeted political messaging.[62]

There is also a noted role that recommender systems (**2.B.1.2**) and targeted advertising (**2.C.1.2**) play in this context; these impacts are discussed later in this paper.

2.A.2.4 Trust: Misinformation

Content moderation can not only suppress diverse viewpoints and contribute to echo chambers and polarisation, but it can also impact truth in news by either removing or unintentionally amplifying misinformation.

The rise of fake news about US federal elections, COVID-19, vaccine safety, and the Russia–Ukraine conflict has highlighted the importance of effective content moderation policies. Platforms have established or strengthened rules regarding false and harmful content distributed via their services.[63] As harmful conspiratorial content grew on platforms, COVID-19 related lockdowns and other preventative measures around the world impacted the content moderation workforce. This led many major platforms to increase their reliance on algorithmic content moderation over human moderators.[64]

However, algorithmic content moderation systems can struggle to differentiate between true and false statements. They find it challenging to understand complex and nuanced speech or identify opinion, parody and satire, and they generally lack understanding about political and cultural contexts. Given platforms' increased reliance on these tools to enforce their policies and terms of service, misinformation can spread because automated systems generate false positive and negatives or fail to remove violating posts while mistakenly

removing acceptable content.[65] This imprecision can limit free speech by silencing dissenting voices.[66]

The role of algorithmic content moderation in elections is a particular concern. Since the 2016 US presidential election, there has been more attention on the spread of misinformation during election campaigns.[67] In March 2021, Twitter's CEO, Jack Dorsey, said during congressional testimony that the site played a role in the storming of the US Capitol building.[68] Media reports suggested an internal Facebook memo on the insurrection noted gaps in Facebook's policies around coordinated authentic harm due to the company's focus on individual violations rather than addressing larger harms across the network at this time.[69] More recently, significant cuts to Twitter's content policy teams have led to concerns about the volume of disinformation about the 2022 run off Brazilian presidential election and the US midterm elections.[70] There is broad concern platforms continue to allow false content that undermines democratic processes.[71]

While misinformation is evident on digital platforms, it remains unclear how much it impacts users' views and behaviour, according to Ofcom.[72] A study published in *Nature* examined exposure to a Russian foreign influence campaign during the 2016 US presidential election, but did not find "evidence of a meaningful relationship" between exposure and attitudes and voting behaviour.[73] Considering our lack of understanding about the true scale or long-term impacts of misinformation, it is crucial to continue to monitor the potential harms and risks associated with automated moderation of platforms.

Misinformation is also relevant to algorithmic recommender systems (**2.B.1.3**) and targeted advertising (**2.C.3.2**). These impacts are discussed later in this paper.

2.A.2.5 Safety: Abuse to marginalised communities impacting participation in online discourse

The transmission of seriously harmful material, such as terrorist and violent extremist content, can have significant impacts on society. It can create and influence radicalised social and political groups that seek to cause fear and anxiety among the community and undermine democratic values. In response to the 2019 terrorist attack on two mosques in Christchurch, New Zealand, the eSafety Commissioner was given powers to request or require an internet service provider to block material that promotes, incites, instructs in or depicts abhorrent violent conduct. These powers seek to minimise the transmission of seriously harmful content and prevent such threats to society.

Recommender systems can also have negative societal impacts by amplifying online content and abuse which targets marginalised communities. This can drive diverse or marginalised voices away from online discourse.[74]

Content moderation systems can also exacerbate structural inequalities. Due to inherent errors in content moderation algorithms, minorities outside the mainstream tend to experience these errors more frequently than others. Algorithms contribute to these issues by filtering out words considered problematic without fully understanding context.[75]

Studies show that platforms remove more content from women and people of colour while failing to remove abuse targeting these groups.[76] Instances of marginalised communities facing disproportionate content reduction and removal are well documented.

Conclusion regarding the potential impact of algorithmic content moderation

In this literature summary, we have examined research on the impacts caused by algorithmic content moderation on users experiencing vulnerability, democratic culture, and the news environment.

The research suggests platforms have significant influence over the information environment. Platforms mediate between governments, businesses and users due to the large volumes of content they oversee. Research also shows that commercial and corporate interests drive the practice of content moderation, and there is a consensus about its potential to inhibit free speech.

Platform influence extends to shaping the circulation and consumption of news through content moderation and recommender systems. While there is limited evidence of echo chambers, filter bubbles and polarisation, it is clear algorithmic content moderation does have some influence over the news environment. It disproportionately impacts those with views and beliefs that are unorthodox or go against accepted societal values and norms.

The strongest indication of harm caused by algorithmic content moderation lies in marginalising opposing views and impacting vulnerable communities facing discrimination. Existing literature reveals racial and gender biases inherent in algorithms themselves. The literature also presents strong, quantitative indications of how algorithmic content moderation affects marginalised groups more. These groups include specific racial and ethnic groups, the LGBTQIA+ community, and people from non-Western countries or those who communicate using languages other than English.

While content moderation has benefits, algorithmic content moderation also poses harms to individuals and society. Algorithmic content moderation is an essential tool to manage the volume of harmful and illegal content on platforms, however it is important to mitigate its harms. Further research to quantify the impacts of algorithmic content moderation, in addition to enhanced transparency and user feedback would be beneficial.

2.B Harms and risks posed by recommender systems

Overview of recommender systems

Recommender systems — also known as content curation systems or ranking algorithms — are used by many digital platforms to prioritise and personalise content for users. This includes social media posts, search results and recommended products. The algorithms analyse user data, such as their interactions on the platform, as well as their contacts' interactions and off-platform activity.[77]

By studying this data, the algorithm can predict how users will react to different types of content, based on factors like type of content, source, engagement by others, and past engagement with similar content. Drawing on this data, recommender algorithms are optimised to achieve various objectives. For example, a service may be seeking to maximise user engagement by maximising time spent on its service (e.g. short-term or long-term engagement,[78] generally to maximise advertising revenue), deliver recommendations that best meet users' needs, or increase the diversity of content that users are exposed to, or some combination of all of these. Consumer responses to content can then be collected and fed back into the system.[79]

This literature summary focuses on consumer-facing digital platform services, and excludes non-consumer systems, such as those that recommend customers to businesses or job applicants to recruiters.

While recommending and ranking content for consumers is not new (e.g. newspapers have long chosen which headlines to put on their front page), algorithmic recommendation differs in several important ways, including with respect to:[80]

- **Data:** Digital platforms collect an unprecedented breadth and depth of data about people and their online behaviours and analyse it in increasingly sophisticated ways;
- **Accuracy and granularity:** Content can be targeted accurately to small groups and even individuals;
- **Iteration:** Online targeting systems continually improve by learning from people's behaviour;
- **Ubiquity:** Content can be distributed at scale and at relatively low cost;[81] and
- **Limited transparency:** Matching people with personalised content limits scrutiny since fewer people see each item and know what others are seeing.

Recommender systems play a significant role in helping consumers sift through the 'information wilderness',[82] which means algorithmic systems construct much of our online activity. [83] For example, about 80% of Netflix viewing hours,[84] 70% of YouTube video views[85] and 35% of purchases on Amazon come from recommendations.[86]

There are different types of recommender systems, and their role can vary across digital platform services.[87] For example, personalised curation is fundamental to social media, but much less important in search, where search terms themselves drive the ranking of results.[88]

Harms and risks

Recommender systems provide substantial benefits to consumers as essential parts of extremely popular digital platform services, but also present a variety of harms and risks.

2.B.1 Individual and societal impacts

2.B.1.1 Fairness: Bias and discrimination

In some circumstances, recommender systems may provide biased recommendations that create risks for individuals and society. These biases can be a result of the algorithms relying on biased training data, which reproduces and amplifies existing biases. This hinders opportunities for particular users, reinforces harmful stereotypes, and facilitates discrimination.[89] For example, as recently as 2016, searching for "three black teenagers" on Google generated results perpetuating racial stereotypes involving violence.[90] Also in 2016, it was uncovered that LinkedIn search results favoured males, potentially resulting in fewer employment opportunities for women.[91]

Discrimination may stem from several sources, such as societal inequalities reflected in the training data or limitations within the data itself.[92] Algorithms can make discrimination worse in ways that are difficult to predict. Algorithms may seem 'objective' but can hide and entrench bias.[93] Concerns that removing protected attributes from consideration (i.e. fairness through unawareness) may not be enough to reduce bias further complicates this challenge.[94]

Issues of bias and discrimination may also arise when algorithms are used for content moderation (**2.A.2.2**) and targeted advertising (**2.C.1.1**).

2.B.1.2 Trust: Echo chambers, filter bubbles and polarisation

Recommender systems inherently prioritise certain content over others. However, there are widespread concerns recommender systems may contribute to echo chambers or filter bubbles by prioritising content that aligns with a user's preferences.[95] This design can limit access to contrary viewpoints and reinforce existing views, potentially leading to societal polarisation over time. While this concern is widely recognised, there is conflicting information about the existence and impact of these phenomena (as described in **2.A.2.3** above).[96] There are also related concerns that echo chambers and filter bubbles could contribute to radicalisation or hinder the news media's ability to identify and scrutinise targeted political messaging.[97]

These concerns extend to issues related to content moderation (**2.A.2.3**) and targeted advertising (**2.C.1.2**).

2.B.1.3 Trust: Misinformation

Recommender systems can inadvertently promote misinformation by prioritising increasingly controversial or shocking content.[98] For example, the MIT Media Lab found false news stories are 70% more likely to be 'retweeted' (shared on Twitter) than true stories, and that true stories take about six times as long to reach 1,500 people as false stories.[99] Digital platforms can also receive monetary benefits from the spread of misinformation on their platforms through additional engagement.[100]

There are widespread concerns about the risks associated with misinformation. At an individual level, consumers may become misinformed about important issues such as vaccinations or political processes, which can have broader implications for society (e.g. stemming the spread of COVID-19). There are further concerns that seeding 'counter-messages' to counter misinformation may be ineffective because it may result in the problematic content being recommended alongside it, thereby increasing exposure.[101]

Digital platforms may enforce content policies such as 'shadowbanning' to slow the spread of misinformation. However, these content policies may themselves rely on algorithms for implementation and enforcement. Moreover, as described in **2.A.2.4,** while misinformation can be seen and spread on digital platform services, the degree to which it influences users' views and behaviour remains unclear. See related discussions about content moderation (**2.A.2.4**) and targeting advertising (**2.C.3.2**).

2.B.1.4 Trust: Impact on news consumption

Using recommender systems on digital platform services can have a variety of impacts on the news that is posted, seen and spread online. These impacts pose potential risks and can affect collective awareness of politics, current affairs and scientific consensus.

For example, in 2014 protests took place in the United States after a police officer fatally shot an African American man named Michael Brown. While this news sparked extensive debate on Twitter, researchers found the topic was "suppressed" on Facebook's algorithmically curated News Feed. Instead, content relating to the 'ice bucket challenge' received more engagement.[102]

Systematic bias in news recommendations — such as promoting or demoting certain types of news or outlets — could potentially influence individual opinions and electoral choices.

This gives those who control recommender systems potential influence over the political process. Ofcom's 2022 review of media plurality and online news noted limited findings about such biases but recognised the need for greater transparency and further research to examine algorithmic bias.[103]

Recommender systems may also shape the type of news that is posted online, since content producers may favour clickbait to increase user engagement.[104] While clickbait can boost short-term engagement with news articles, it can erode perceptions of credibility and quality in news media overall.[105]

2.B.1.5 Trust and safety: Inauthentic account use

Third parties, including malicious actors using networks of inauthentic accounts, can exploit online targeting systems. They do this by artificially inflating views, likes, shares and other metrics to manipulate content recommender systems. This manipulation increases the chances of content promoted by inauthentic accounts being recommended to a wider audience.[106]

Companies and governments may have political and economic incentives to engage in such behaviour. These incentives include amplifying support for their decisions, silencing criticism, or fuelling division among others.[107] There are concerns about how cheap it can be to carry out such actions. For example, researchers were able to buy more than 3,000 comments, 25,000 likes, 20,000 views and 5,000 followers across platforms such as Instagram, Facebook, Twitter and YouTube for just €300.[108] Researchers at the Oxford Internet Institute found that organised campaigns involving political manipulation on social media occurred in 81 countries in 2020, including campaigns organised by state actors.[109]

Similarly, users could try to undermine or circumvent moderation systems.

2.B.1.6 Safety: Spread of terrorist and extreme violence content

Recommender systems may also allow the spread of terrorist and extreme violence content online before it is taken down by digital platforms. This content can instil fear and anxiety in individuals who see it, and create a sense of unease and insecurity among the general public.

Livestreamed acts of violent extremism, such as the 2019 terrorist attack on two mosques in Christchurch, exemplify how perpetrators can reach global audiences through smaller platforms, causing widespread harm to unsuspecting users. The New Zealand Government's 2019 Report of the Royal Commission of Inquiry into the terrorist attack on the Christchurch mosques outlined how online services are a key platform for terrorist radicalisation and recruitment, and for developing and sharing extreme right-wing views.[110]

2.B.1.7 Safety: Normalising harmful content

Recommender systems have the potential to amplify harmful and extreme content, which can have significant individual impacts on those exposed to such material. On a broader societal level, the amplification of discriminatory content promoting sexism, misogyny, homophobia or racism, can normalise prejudice or hate. This may also contribute to radicalisation towards terrorism and violent extremism and undermine societal values and cohesion.

For children specifically, recommender systems can promote unrealistic 'ideals' of body types and reinforce beauty stereotypes. It may also normalise the sexualisation of young people. This includes previous reports of sexualised content being recommended to children.[111]

Reports have shown platforms recommending content related to body image issues and disordered eating. Investigations by Reset Australia revealed that Instagram amplifies pro-anorexia content to teens and children as young at 10. It reported that 32% of girls see content that promotes, glamorises or normalises extreme weight loss several times a day.[112]

Identifying harmful content from recommender systems can be difficult due to various factors. The experience of harm differs depending on individual circumstances or communities involved. Whether certain content is harmful typically depends on the specific context, such as the identity or other relevant factors pertaining to the person who encounters it.

2.B.2 Individual impacts

2.B.2.1 Safety: Users experiencing vulnerability

Like content moderation algorithms (**2.A.1.1**), recommender systems can also pose safety risks for users experiencing vulnerability. Some examples of harm caused or exacerbated by recommender systems include:

- **Children's exposure to inappropriate content:** Children may be exposed to adult content or face higher risks of online sexual exploitation due to friend or follower suggestions that pressure them to interact with dangerous adults.[113]
- **Exploiting or exacerbating physical/psychological disorders:** Recommender systems may serve increasingly extreme content to someone because they have viewed similar material.[114] This may impact individuals in various ways, such as increasing their exposure to content promoting self-harm,[115] cyberbullying content or beauty stereotypes.[116]
- **Exploiting or excessive internet use:** There are concerns about recommender systems designed to maximise engagement and clicks. Certain groups, such as children, older people, people with learning disabilities, and people with addictions may be especially vulnerable.[117] Excessive internet use refers to a state where individuals "lose control" and continue using the internet despite experiencing negative outcomes. Some studies suggest recommender systems drive excessive usage of video websites.[118]
- **Manipulating consumer choice:** When designed or calibrated inappropriately, recommender algorithms can (inadvertently or intentionally) mislead consumers about the relative costs or benefits of products and services. For example, in January 2020 the Federal Court of Australia found Trivago had misled consumers in representing that its website would help users find the best or cheapest hotel deals available. Instead, Trivago's recommender algorithm had been designed to favour hotel booking sites that paid Trivago the highest referral fees.[119]

Preventing specific instances of harm (and measuring the severity of harm) is challenging because certain content may harm some users but not others.[120] The personalised and dynamic nature of the online space allows platforms using recommender systems to systematically present users with choice architectures that "can be specifically designed to exploit each individual user's particular vulnerabilities."[121] See also section **2.C.2.2** on targeted advertising.

2.B.2.2 Safety: Exposure to harmful content

Recommender systems can contribute to the spread of harmful content. When these systems are designed to prioritise human engagement, they run the risk of exploiting human cognitive biases, drawing people to shocking and extreme content. Recommender systems can end up pushing content that is harmful but not illegal. This includes 'borderline' content that is not detected by moderation practices, yet may violate terms of service, community standards or local law. Recommender systems may also serve increasingly extreme content to someone because they have previously viewed similar material, which raises concerns about the association between recommender systems and radicalisation.[122] See also section **2.A.1.2** above on content moderation.

2.B.2.3 Fairness: Incentivising increased collection or storage of data

Algorithms use and analyse user information to build profiles that help with targeting systems, such as recommender systems and targeted advertising. Detailed profiles may improve the accuracy and impact of recommendations and advertising, which incentivises platforms to collect large amounts of personal information for their own commercial interests and to influence consumer choices.[123] While collecting more user data can improve the quality of services, it also increases the risk of users losing control over increasing amounts of their personal information. A survey conducted by the OAIC in 2020 found that more than 50% of Australians are uncomfortable with online businesses and digital platforms keeping information on what they have said and done online.[124] The survey also found that 81% of Australians believe asking for personal information that does not seem relevant for the purpose of the transaction is a misuse.[125]

2.B.3 Societal impacts

2.B.3.1 Fairness: Threats to competition

Firms may use recommender systems to engage in exclusionary conduct, which means restricting or undermining rivals' ability to compete. This allows them to maintain or advance their own position in the market. For example, these systems can be used for anti-competitive self-preferencing by platform services that include their own offerings alongside competing third-party offerings in search results.[126] For example, the European Commission found that Google abused its dominant market position in online search to give preferential treatment to its own comparison-shopping service.[127] Exclusionary conduct can reduce the incentive for new and existing firms to develop and provide improved services, and subsequently reduce consumer choice.

Conclusion regarding the potential impact of algorithms in recommender systems

Recommender systems are an essential part of many popular digital platform services. However, their use has given rise to a wide range of individual and societal risks. These risks include negative impacts on public discourse, safety, privacy, fairness for individuals, and competition in relevant markets.

These are generally not the intended effects of recommender systems. Their main goal is to help users find relevant and desirable content. However, the research surveyed in this paper suggests many of these unintended effects are made possible by:

- A lack of transparency around the design and operation of recommender systems and the data they are trained on;

- An emphasis on highly personalised recommendations; and
- A handful of powerful digital platforms having significant control over the content users are exposed to online.

The literature suggests that digital platform firms, regulators and researchers should consider these issues in seeking to address the harms and risks recommender systems pose. In the future design and regulation of services that use these systems, it may be necessary to find a balance between the benefits they provide to users and the risks and harms they present.

2.C Harms and risks posed by algorithms used in targeted advertising

Overview of targeted advertising systems

Targeted advertising, also known as behavioural targeting, refers to advertising "targeted to individual consumers based on inferences about their personal attributes, such as their interests, demographics or characteristics".[128] Unlike traditional forms of advertising, such as in a physical newspaper, targeted advertising allows different users viewing the same content on the same website to see different advertisements.[129]

Targeted advertisements can take a variety of forms, including display advertisements, video advertisements and sponsored content.[130] They can promote diverse subjects, including goods and services, job opportunities and campaigns to promote a perspective, such as a political view or health belief.

Algorithms play a crucial role in the targeted advertising process, such as:

- Composing the audience for the advertisement, including through identifying lookalike audiences (other users who share similar characteristics to users in an audience provided by the advertiser);
- Determining how the advertisement is presented, such as its design;
- Making buying and bidding decisions for advertisers;
- Determining which advertiser's bid should win the auction for an advertising opportunity; and
- Providing diagnostic data on the effectiveness of advertising.[131]

To carry out these tasks effectively, algorithms rely heavily on detailed user data. Algorithms use this data to build user profiles for targeting purposes and pricing strategies. Detailed user data also helps advertisers measure the performance of their ads and enables algorithms to learn from real-time user behaviour for more effective future advertising.[132] While targeting can be beneficial, by providing relevant or interesting advertising to users, commentators have raised concerns about potential harms to consumers and society as detailed below.

2.C.1 Individual and societal impacts

2.C.1.1 Fairness: Bias and discrimination

Algorithms used in targeted advertising on social media can skew the delivery of advertisements, with certain demographic groups being underrepresented or excluded from certain advertisements.[133] These biases in automated advertising algorithms can lead to discrimination based on factors such as race and gender, which restricts awareness of

opportunities such as job openings for historically disadvantaged groups.[134] These consequences entrench existing patterns of social exclusion and economic inequality.[135]

The systems behind targeted advertising can be used to discriminately target certain demographics through the selection of certain audiences during ad creation.[136] Even if explicitly discriminatory attributes are removed, discriminatory targeting can persist through the use of proxy characteristics.[137]

Moreover, even when a neutral audience is selected, advertisement delivery algorithms can introduce bias. Platforms show 'relevant' advertisements based on individual user profiles and 'lookalike groups', which may skew advertisement delivery to a subgroup of an advertiser's intended audience.[138] In addition, pricing and bidding algorithms can contribute to gender-based biases because displaying advertisements to women tends to be more expensive than showing them to men.[139] This means algorithms may skew advertisements towards an audience in ways that were not intended by advertisers. Similar issues can arise with content moderation (**2.A.2.2**) and recommender systems (**2.A.2.2**).

2.C.1.2 Trust: Echo chambers, filter bubbles and polarisation

Algorithms in targeted advertising on social media can distribute advertisements unfairly, where certain demographic groups are not adequately represented or left out. As a result, individuals may only see certain types of advertisements, limiting the range of content they are exposed to.[140] This repetitive display and reinforcement of certain values and attitudes can create a filter bubble or echo chamber effect. It restricts an individual's exposure to diverse perspectives and ideas by narrowing their online experiences.[141] See also section **2.A.2.3** (content moderation) and **2.B.1.2** (recommender systems).

2.C.2 Individual impacts

2.C.2.1 Fairness, Trust and Safety: Lack of individual control of personal information

One of the negative consequences of incentivised data collection and storage is that individuals have limited control over their personal information. Studies show that it is not always clear to consumers what personal information is being collected and how it is shared, such as through bid requests in real-time bidding auctions.[142] Complex terms and conditions purport to gain consumer consent even though consumers may not understand them.[143] As a result, consumers are left with little control over how their personal information is used and who it is shared with. See also section **2.B.2.3** above (recommender systems).

2.C.2.2 Fairness and safety: Users experiencing vulnerability

Targeted advertising has the potential to manipulate consumer preferences and exploit individuals experiencing vulnerability. Advertisers can use data and algorithms to identify consumers who are susceptible to certain types of advertising. For example, advertisers may target individuals with low self-esteem by showing them tailored advertisements for products such as diet or cosmetic items. Advertisers might also direct gambling advertisements to children or frequent gamblers.[144] In addition, A/B testing can be employed to determine which version of an advertisement has the greatest impact on an individual user.[145] This testing involves randomly displaying different versions of an advertisement simultaneously to various visitors. Such practices can be particularly harmful when users experiencing vulnerability are targeted with advertisements for products that may not serve their best interests — or with scam advertisements — as noted in the recommender systems section.[146] As described above, there are related risks associated with using algorithms for content moderation (**2.A.1.1**) and recommender systems (**2.B.2.1**).

2.C.3 Societal impacts

2.C.3.1 Trust: Restricting transparency of digital platform activities

Using algorithms to personalise advertising limits transparency in targeted advertising on platforms. Instead of being viewed by the general public, separate advertising is shown to each individual.[147] While platforms frequently provide advertising transparency dashboards, these dashboards only offer basic insights and fail to provide a complete understanding of how platforms implement targeted advertising.[148] These dashboards may aggregate or abstract important information, remove historical data, and obscure detailed data necessary to identify patterns in reach and targeting that might indicate discrimination or predatory advertising. Additionally, advertising transparency dashboards are not independently verified. As a result, it is difficult to know what advertisements individuals are receiving and why.

The lack of transparency regarding how information is used for targeting and which advertising individuals receive undermines public accountability for the advertisements being disseminated and their impacts on recipients, such as discrimination.[149] It also makes it more difficult for individuals to recognise when they are the subject of discrimination or manipulation, because they cannot compare the advertisements they receive to those received by others.[150]

2.C.3.2 Trust: Misinformation

Advertising that focuses on promoting certain concepts or ideas (rather than tangible goods, services or opportunities) can cause additional, context-specific harms to society through the broader effects of polarisation, misinformation and disinformation. Concerns have been raised in relation to various topics, including political messaging and health risks related to anti-vaccination material.[151] To illustrate this point further, we will discuss the specific harms associated with targeted political advertising as identified in the literature.

Impacts on political process

Targeted advertising can be used in the political process to ‘micro-target’ individuals based on their location or political opinions. It aims to influence their political engagement through tailored messaging.[152] While political advertising can inform people about political groups and policies, algorithms can be used to identify existing tensions, resentments and anxieties. These factors can be subtly exploited to manipulate individuals’ political behaviour, drive polarisation, and contribute to the development of harmful echo chambers.[153]

The use of algorithms in targeted political advertising distinguishes it from other political advertising. By delivering customised messages to different groups it allows for broader reach while minimising the risk of backlash.[154] In addition, using algorithms helps to improve messaging by testing and measuring engagement.[155] Finally, there are reports that advertising algorithms favour polarising messaging because they assign lower costs to content more likely to generate user engagement.[156] The limited transparency of advertising delivery and the practice of displaying advertisements to individuals who likely agree with their substance leads to limited public oversight of political advertising.[157] Related concerns may also arise where influencers are used to create content for political advertising.[158] See also section **2.A.2.4** (content moderation) and **2.B.1.3** (recommender systems).

Conclusion regarding potential impacts of algorithms in targeted advertising

Algorithms play various roles in targeted advertising. This includes determining the target audience, selecting how advertisements are presented, and setting prices and winning bids in the sale of advertising space.

The research we reviewed identified a range of harms that can result from targeted advertising at both the individual and societal levels. Some of these harms, such as bias and discrimination, appear to stem from the way targeting is conducted or how algorithms operate. This means they can arise from intentional choices, or unintentionally. Other negative effects flow from using profiles to target advertising, which can incentivise increased data collection and promote echo chambers. The nature of delivering personalised advertising also affects transparency. Participants have limited visibility into related algorithms, which can contribute to concerns about misinformation. It is crucial for digital platforms, regulators and policy departments to be aware of these potential impacts and consider how to mitigate them. Some research suggests a need for more Australian data to understand the local impact of targeted advertising.

2.D Summary of harms and risks identified

Error! Not a valid bookmark self-reference. below summarises the harms and risks identified in Section 2.

Table 1: Summary of harms and risks

Content moderation			
Level of Harm	*Fairness*	*Trust*	*Safety*
Individual			Users experiencing vulnerability
			Exposure to harmful content
Societal	Participation in society and promotion of democratic values	Echo chambers, filter bubbles and polarisation	Threats to society
	Bias and discrimination	Misinformation	Abuse to marginalised communities impacting participation in online discourse
Recommender systems			
Level of Harm	*Fairness*	*Trust*	*Safety*
Individual	Bias and discrimination	Echo chambers, filter bubbles and polarisation	Users experiencing vulnerability
	Incentivising increased collection or storage of data	Impact on news consumption	Exposure to harmful content
		Inauthentic account use	Inauthentic account use
		Misinformation	Spread of terrorist and extreme violence content
			Normalising harmful content
Societal	Bias and discrimination	Echo chambers, filter bubbles and polarisation	Inauthentic account use
	Threats to competition	Misinformation	Spread of terrorist and extreme violence content
		Impact on news consumption	Abuse to marginalised communities impacting participation in online discourse
		Inauthentic account use	Normalising harmful content
Targeted advertising			
Level of Harm	*Fairness*	*Trust*	*Safety*
Individual	Bias and discrimination	Echo chambers, filter bubbles and polarisation	Lack of individual control of personal information
	Lack of individual control of personal information	Lack of individual control of personal information	Users experiencing vulnerability
	Users experiencing vulnerability		
Societal	Bias and discrimination	Restricting transparency of digital platform activities	
		Echo chambers, filter bubbles and polarisation	
		Misinformation	

3 Regulatory initiatives to address harms and risks posed by algorithms

Annex 1 provides some relevant examples of proposed or enacted regulatory initiatives aimed at addressing the harms and risks posed by algorithms, both domestically and overseas. This list is not exhaustive, nor does it evaluate these initiatives.

Transparency initiatives are being developed and implemented around the world to address online harms. For example, the Digital Services Act[159] (DSA) in the European Union requires digital platforms (online platforms) to be transparent about targeted advertising, including details about how advertisements are targeted. In other cases, digital platforms may voluntarily provide broader transparency about their efforts to address harms occurring on their platforms through initiatives such as regular reporting under the European Union's Strengthened Code of Practice on Disinformation.[160] Some countries have specific regulatory requirements for reporting against safety expectations related to online issues, such as the Online Safety Act's Basic Online Safety Expectations[161] in Australia, and the European Union's Digital Services Act transparency requirements for safety issues, including recommender systems).

Certain jurisdictions will require regulated entities to empower end users by providing choices. An example is China's law on recommender systems.[162] Further, some regulatory initiatives will prohibit digital platforms from engaging in certain conduct to address risks, as in the case of the Digital Markets Act[163] (DMA) in the European Union which will prohibit digital platforms from engaging in self-preferencing.

In addition to regulatory requirements, there are ongoing efforts to establish best practices to guide industry in ethical AI use. For example, the Australian government[164], the OECD[165], the European Commission[166] and UNESCO[167], among others' have developed ethical principles for AI usage. Additionally, forums such as the United Kingdom's Digital Regulators Co-operation Forum foster deeper cooperation among regulators to deal with emerging harms effectively.

4 Conclusion

This literature summary has identified a wide variety of harms and risks associated with algorithms in the areas of content moderation, recommender systems and targeted advertising. Several harms and risks were common to all three types of algorithms:

- Replicating existing societal bias and discrimination;
- Distributing misinformation; and
- Presenting particular harms to users experiencing vulnerability.

Analysing these individual and societal risks, along with others noted in this paper, demonstrates the serious impacts the design and mechanics of these systems can have on digital platform users. These algorithms often operate invisibly to users and are not fully transparent to researchers or regulators. This suggests digital platforms have substantial influence that is worthy of further scrutiny. As a result, governments and regulators around the world, including those in Australia, are increasingly considering and implementing new initiatives to manage the risks posed by platforms' algorithms, as is summarised in **Annex 1**.

In terms of future directions for research, we note that much of the literature reviewed in this paper focused on studies conducted in the United States, the United Kingdom and Europe.

There is limited research exploring how these harms and risks affect Australians. Also, while many potential harms and risks have been identified, there may be mixed information regarding their extent such as with polarisation or misinformation.

By conducting this literature summary, DP-REG members have gained a shared understanding of the harms and risks associated with common types of algorithms. This deeper knowledge will support the future work of individual regulators and of DP-REG. We hope that this document will provide a valuable reference for regulators as they continue to monitor and contribute to domestic policy development relevant to our regulatory responsibilities. This includes the Online Safety Act, the Basic Online Safety Expectations, and the next steps following the Attorney-General's Department's Review of the Privacy Act.

Annex 1: Examples of regulatory initiatives to address algorithmic risks

Regulatory requirements	Country/ Region	Example of regulatory initiative	Relevant category of harm	Harms addressed
Transparency; accountability	AU	Online Safety Act 2021 Online Safety (Basic Online Safety Expectations) Determination 2022	content moderation Recommender Systems	Illegal and harmful content and activity
Transparency	AU	Australian Code of Practice on Disinformation and Misinformation	All	Misinformation
Transparency	China	Internet Information Service Algorithmic Recommendation Management Provisions	Recommender systems	Restricting transparency of digital platform activities'
Transparency	EU	Digital Markets Act – transparency about pricing of advertising services	Targeted advertising Recommender systems	Restricting transparency of digital platform activities' Threats to competition
Transparency	EU	Digital Services Act – article 27 – transparency on the main parameters used in recommendation rankings	Recommender systems	Restricting transparency of digital platform activities'
Transparency	EU	Digital Services Act article 26(1) – online platform providers must provide certain information for each ad presented to the individual, including the parameters used to determine who the ad is presented to	Targeted advertising	Lack of individual control of personal information Exploiting or exacerbating physical/psychological disorders Bias and discrimination Restricting/eliminating/manipulating user choice Restricting transparency of digital platform activities
Transparency	EU	Digital Services Act article 39 – very large online platforms must have publicly available ad repositories that meet certain criteria	Targeted advertising	Lack of individual control of personal information Exploiting or exacerbating physical/psychological disorder Bias and discrimination Restricting/eliminating/manipulating user choice Restricting transparency of digital platform activities
Transparency	EU	Digital Services Act article 46 – provides for the development of voluntary codes of conduct for online advertising	Targeted advertising	Lack of individual control of personal information Exploiting or exacerbating physical/psychological disorders Bias and discrimination Restricting/eliminating/manipulating user choice

Regulatory requirements	Country/ Region	Example of regulatory initiative	Relevant category of harm	Harms addressed
				Restricting transparency of digital platform activities
Transparency	EU	Strengthened Code of Practice on Disinformation 2022	Content moderation, targeted advertising, recommender systems	Misinformation
Transparency	EU	European Centre for Algorithmic Transparency (ECAT)	Recommender systems	Assess risks stemming from algorithmic systems
Transparency	UK	Online Safety Bill (Schedule 8 – Transparency reports)	Recommender systems, content moderation	Illegal and harmful content and activity
Transparency	France	Les enjeux de la loi contre la manipulation de l'information (Laws against the manipulation of information)	Recommender systems, content moderation. advertising	Misinformation
Minimum standards for industry	AU	Online Safety Act 2021: Online safety industry codes	All	Illegal and harmful content
Privacy safeguards	AU	The Attorney-General's Department's final report in the Review of the Privacy Act 1988 makes 116 proposals for privacy reform, including proposals that are relevant to algorithms using personal information and specific proposals relating to direct marketing and targeting.	All	Incentivising increased collection and storage of personal information Users experiencing vulnerability Bias and discrimination Restricting/eliminating/manipulating user choice Restricting transparency of digital platform activities Lack of individual control of personal information
Safeguards	EU	Artificial Intelligence Act	Content moderation; recommender systems	Exposure to harmful content and behaviour
Principles and best practices for industry; transparency	Global	OECD's Principles of AI	All	Security; safety

Regulatory requirements	Country/ Region	Example of regulatory initiative	Relevant category of harm	Harms addressed
Principles and best practices for industry; transparency	AU	Australia's AI ethics principles	All	Security; safety
Principles and best practices for industry; transparency	EU	European Commission Ethics Guidelines For Trustworthy AI	All	Security; safety
Principles and best practices for industry; transparency	Global	UNESCO Recommendation on the Ethics of Artificial Intelligence	All	Security; safety
Australian Government collaboration	AU	DISR's Digital Economy Branch	Content moderation; recommender systems	Exposure to harmful content; security; safety
Regulatory collaboration and coordination	AU	DP-REG	All	All
Regulatory collaboration and coordination	UK	Digital Regulation Cooperation Forum	All	All
Regulatory collaboration and coordination	Global	Global Online Safety Regulators Network	All	Safety
Empower user choice	EU	Digital Services Act – article 38 – a choice to opt-out of use of use of profiling in recommender systems	Recommender systems	User choice over targeting
Empower user choice	China	Internet Information Service Algorithmic Recommendation Management Provisions	Recommender systems	User choice over targeting
Conduct obligations	EU	Digital Markets Act – prohibition of self-preferencing	Recommender systems	Threats to competition (self-preferencing)
Conduct obligations	EU	Digital Services Act risk assessment and mitigation	Recommender systems	Bias and discrimination

Regulatory requirements	Country/ Region	Example of regulatory initiative	Relevant category of harm	Harms addressed
Conduct obligations	UK	Digital Markets Bill – potential prohibition of anti-competitive self-preferencing	Recommender systems Targeted advertising	Threats to competition
Conduct obligations	AU	ACCC Regulatory reform report – potential prohibition of anti-competitive self-preferencing	Recommender systems Targeted Advertising	Threat to competition
Conduct obligations	EU	Digital Services Act articles 26(3) and 28(2) - prohibits advertisement based on profiling using the data of children or special categories of personal information	Targeted advertising	Incentivising increased collection or storage of data Lack of individual control of personal information Users experiencing vulnerability Bias and discrimination
Anti-discrimination	India	Consumer Protection (eCommerce) Rules 2020 s 4(11) – prohibits e-commerce entities from manipulating price offerings or discriminating between consumers based on arbitrary classifications.	Targeted advertising	Price discrimination
Anti-discrimination	USA	Various legislation including Title VII of the Civil Rights Act 1991, Age Discrimination in Employment Act 1967, Fair Housing Act 1968, Equal Credit Opportunity Act 1974, which have influenced changes in platforms' policies.[168]	Targeted Advertising	Bias and discrimination
Opt-out	USA – California	California Consumer Privacy Act 2018 s 1798.135 – businesses using or disclosing personal information for purposes other than what is necessary to provide goods or services must incorporate a 'Do Not Sell My Personal Information' link on their website home page, which takes consumers to a designated webpage where they may 'opt-out'	Targeted Advertising	Incentivising increased collection or storage of data Lack of individual control of personal information Users experiencing vulnerability Bias and discrimination
Opt-out	AU	Privacy Act Review Report 2022 – includes a proposal to provide individuals with an unqualified right to opt-out of receiving targeted advertising	Targeted Advertising	Incentivising increased collection or storage of data Lack of individual control of personal information Users experiencing vulnerability Bias and discrimination

Endnotes

[1] Autorite de la Concurrence & Bundeskartellamt, Algorithms and Competition, November 2019.

[2] Centre for Data Ethics and Innovation, Review into bias in algorithmic decision making, November 2020, accessed 11 January 2023.

[3] DP-REG's infographic notes that '*Each regulator brings an important and distinct lens based on their different remits to intersecting issues on digital platforms. Collaboration and coordination on these topics ensures a proportionate response and shared focus on improving our digital economy by making it a* ***safe, trusted, fair****, innovative and competitive space*'. For the purposes of this literature summary, threats to competition are included under the broader theme of fairness.

[4] For example, Facebook's recommendation system tries to identify and reduce the prominence of posts with exaggerated or sensational health claims. In this case, its system identifies commonly used phrases to predict which posts are likely to breach their policies and refers them to human moderators or fact checkers, who can then decide to downrank it.

Facebook, Addressing Sensational Health Claims, 2 July 2019, accessed 10 January 2023.

[5] Though, 'shadowbanning' may raise risks of its own, as it may disproportionately impact content creators who deal with sensitive topics (e.g. providing education about sex or self-harm) or may occur suddenly without explanation with limited options for review. eSafety Commissioner, Position statement - Recommender systems and algorithms, December 2022.

[6] R. Gorwa, R. Binns, and C. Katzenbach, Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7:1, 2020.

Recommender systems more generally are explored in section 2.B. While similar to softer moderation measures, this literature summary distinguishes recommender systems which are used to prioritise and personalise content based on user data to optimise for a desired outcome from softer moderation measures which are targeted to specific content but not personalised to users.

[7] E. Douek, Governing Online Speech: From 'posts-as-Trumps' to Proportionality and Probability, *Columbia Law Review*, 121:3 (2021), p. 759–833.

[8] D. Thakur, and E. Llansó, Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis, Center for Democracy & Technology, 20 May 2021.

[9] D. Thakur, and E. Llansó, Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis, Center for Democracy & Technology, 20 May 2021.

[10] PhotoDNA, one of the most widely used hash matching tools has a reported error rate of 1 in 50 billion. F, Hany, Statement to the House Committee on Energy and Commerce Hearing: Fostering a Healthier Internet to Protect Consumers, 16 October 2019.

[11] All tech is Human, *AI and Human Rights*, 27 June 2022, accessed 11 January 2023.

[12] V. Gill, L. Monk, L. Day, Qualitative research project to investigated the impact of online harm on children, April 2022, accessed 11 January; eSafety Commissioner, Online hate speech: Findings from Australia, New Zealand and Europe, January 2020, accessed 12 January 2023.

[13] eSafety Commissioner, Mind the Gap: Parental awareness of children's exposure to risks online, February 2020.

[14] ACCC, Targeting scams: Report of the ACCC on scams activity 2021, July 2022, accessed 10 February 2023.

[15] ACCC, ACCC takes action over alleged misleading conduct by Meta for publishing scam celebrity crypto ads on Facebook, 18 March 2022, accessed 18 January 2023.

[16] D. Konikoff, Gatekeepers of toxicity: Reconceptualizing Twitter's abuse and hate speech policies, *Policy & Internet*, 13:4 (2021), p. 502-521.

[17] Z. Reeve, Repeated and Extensive Exposure to Online Terrorist Content: Counter-Terrorism Internet Referral Unit Perceived Stresses and Strategies, Studies in Conflict & Terrorism, 46:6 (2023), p. 888-912.

[18] eSafety Commissioner, Mind the Gap: Parental awareness of children's exposure to risks online, February 2020, accessed 12 January 2023.

[19] For further detail of what constitutes R18+ and X18+ content, see eSafety Commissioner, Online Content Scheme Regulatory Guidance, December 2021, p. 4.

[20] V. Jaynes, and I. Wick, Risky by Design: Recommendation Systems, 5Rights Foundation, 2022, accessed 10 January 2023.

[21] DataReportal, 2023 Global Digital Report, 26 January 2023, accessed 1 February 2023.

[22] G. De Gregorio, 'Democratising online content moderation: A constitutional framework,' Computer Law & Security Review, 36, 2020; On 'freedom of expression', we recognise that the First Amendment protects speech from government interference. As such, platforms are entities who can determine their own policies on speech rights. However, we do note that the perceived right to a freedom of expression permeates broader society and influences the way American users judge how platforms apply content moderation policies to content posted on their services.

[23] K. Klonick, The New Governors: The People, Rules, and Processes Governing Online Speech, Harv. L. Rev, 131:6 (2018), p.1598-1670.

[24] G. De Gregorio, 'Democratising online content moderation: A constitutional framework,' Computer Law & Security Review, 36, 2020.

[25] R. Gorwa, R. Binns, and C. Katzenbach, Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7:1, 2020.

[26] C. Katzenbach, and L. Ulbricht, Algorithmic governance, *Internet Policy Review*, 8:4 (2019).

[27] K. Klonick, The New Governors: The People, Rules, and Processes Governing Online Speech, Harv. L. Rev, 131:6 (2018), p.1598-1670.

[28] R. Caplan, Content or Context Moderation? Artisanal, community-reliant, and industrial approaches, Data & Society Research Institute, 14 November 2018.

[29] N. P. Suzor, S. M. West, A. Quodling, and J. York, What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation, *International Journal of Communication,* 13 (2019), p. 1526-1543.

[30] P. Waldron, One-size-fits-all content moderation fails the Global South, Cornell Chronicle, 13 April 2023, accessed 15 May 2023.

[31] BBC, Lese-majeste explained: How Thailand forbids insult of its royalty, 6 November 2017, accessed 15 May 2023.

[32] European Parliament Research Service, Holocaust denial in criminal law: Legal frameworks in selected EU member states, January 2022.

[33] K. Klonick, The New Governors: The People, Rules, and Processes Governing Online Speech, Harv. L. Rev, 131:6 (2018), p.1664

[34] Ibid, 1653.

[35] T. Lorenz, *Internet 'algospeak' is changing our language in real time, from 'nip nops' to 'le dollar bean*, The Washington Post, 8 April 2022, accessed 17 January 2023.

[36] D. Thakur, and E. Llansó, Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis, Center for Democracy & Technology, 20 May 2021.

[37] VentureBeat, Report: Computer vision teams worldwide say projects are delayed by insufficient data, 24 December 2021, accessed 15 May 2023.

As for the issue of quality of visual data, for example, single sentence descriptions to label images may be insufficient to describe the complex scene depicted by an image. See S. Frolov, T. Hinx, F.

Raue, J. Hees and A. Dengel, Adversarial text-to-image synthesis: A review, Neural Networks, 144 (2021), p. 201.

[38] E. Pirkova, M. Kettemann, M. Wisniak, M. Scheinin, E. Bevensee, K. Pentney, L. Woods, L. Heitz, B. Kostic, K. Rozgonyi, H. Sargeant, J. Haas, and V. Joler, Spotlight on Artificial Intelligence and Freedom of Expression A Policy Manual, Office of the Representative on Freedom of the Media Organization for Security and Co-operation in Europe, 2021; D. Thakur, and E. Llansó, Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis, Center for Democracy & Technology, 20 May 2021, accessed 10 January 2023; T. Gillespie, Content moderation, AI, and the question of scale, *Big Data & Society*, 7:2 (2020).

[39] D. Thakur, and E. Llansó, Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis, Center for Democracy & Technology, 20 May 2021, accessed 10 January 2023; K. Wiggers, Foundation models risk exacerbating ML's ethical challenges, Venture Beat, 18 August 2021, accessed 10 January 2023.

[40] S. Singh, Everything in Moderation, New America, 15 July 2019, accessed 18 January 2023.

[41] J. Cobbe, Algorithmic Censorship by Social Platforms: Power and Resistance, *Philosophy & Technology*, 34 (2021), p. 739-776.

[42] R. O'Kane, Meta's Private Speech Governance and the Role of the Oversight Board: Lessons from the Board's First Decisions, *25 Stanford Technology Law Review*, 25:2 (2021), p.167-209.

[43] S. Ghaffary, The Algorithms That Detect Hate Speech Online Are Biased against Black People, Vox, 15 August 2019 accessed 10 January 2023; M. Sap, D. Card, S. Gabriel, Y. Choi and N.A. Smith, The Risk of Racial Bias in Hate Speech Detection, In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 1668–78.

[44] C. Sinders, Toxicity and tone are not the same thing: Analyzing the new Google API on toxicity, Perspective API. Medium, 24 February 2017, accessed 9 January 2023.

[45] F. Ryan, A. Fritz and D. Impiombato, TikTok and WeChat: curating and controlling global information flows, International Cyber Policy Centre, ASPI Policy Brief, 2020, accessed 8 February 2023.

[46] F. Ryan, A. Fritz and D. Impiombato, TikTok and WeChat: curating and controlling global information flows, International Cyber Policy Centre, ASPI Policy Brief, 2020, accessed 8 February 2023.

[47] S.T. Roberts, Behind the Screen: Content Moderation in the Shadows of Social Media. New Haven: Yale University Press, 2019.

[48] For example, 34% of Facebook news users say more than half of the feed is from news sources. The figure is from a survey of 337, 206 adult Australians who participated in an online questionnaire for the Digital News Report: Australia 2022. See S. Park, K. Mcguinness, C. Fisher, JY. LEE, K. Mccallum and D. Nolan, Digital News Report: Australia 2022, News Media Research Centre, University of Canberra, 2022, accessed 8 February 2023.

[49] K. Klonick, The New Governors: The People, Rules, and Processes Governing Online Speech, Harv. L. Rev, 131:6 (2018), p. 1666.

[50] Ibid.

[51] An echo chamber is used to describe "a bounded, enclosed media space that has the potential to both magnify the messages delivered within it and insulate them from rebuttal." See K.H. Jamieson, & J.N. Cappella, Echo chamber: Rush Limbaugh and the conservative media establishment, Oxford and New York: Oxford University Press, 2008.

[52] "An echo chamber is a form of bubble... [it is] is an echo chamber primarily produced by ranking algorithms engaged in passive personalisation without any active choice on our part, a possible outcome of specific aspects of how news and information is distributed online." See: R.A. Arguedas, C. T. Robertson, R. Fletcher and R.K. Nielsen, Echo Chambers, Filter Bubbles, and Polarisation: a Literature Review, Reuters Institute for the Study of Journalism, 19 January 2022, accessed 13 January 2023.

[53] R.A. Arguedas, C. T. Robertson, R. Fletcher and R.K. Nielsen, Echo Chambers, Filter Bubbles, and Polarisation: a Literature Review, Reuters Institute for the Study of Journalism, 13 January 2022, accessed 13 January 2023.

[54] The Oxford Internet Institute's Literature Review offers a comprehensive assessment on these effects: R.A. Arguedas, C. T. Robertson, R. Fletcher and R.K. Nielsen, Echo Chambers, Filter Bubbles, and Polarisation: a Literature Review, Reuters Institute for the Study of Journalism, 19 January 2022, accessed 13 January 2023.

[55] Ibid.

[56] Ibid.

[57] Ideological polarisation refers to the degree to which people disagree about political issues. From: R.A. Arguedas, C. T. Robertson, R. Fletcher and R.K. Nielsen, Echo Chambers, Filter Bubbles, and Polarisation: a Literature Review, Reuters Institute for the Study of Journalism, 19 January 2022, accessed 13 January 2023.

[58] Affective polarisation refers to people's feelings about the 'other side' – those they disagree with on a given issue. From: R.A. Arguedas, C. T. Robertson, R. Fletcher and R.K. Nielsen, Echo Chambers, Filter Bubbles, and Polarisation: a Literature Review, Reuters Institute for the Study of Journalism, 19 January 2022, accessed 13 January 2023.

[59] News audience polarisation refers to the degree to which audiences for news outlets in a given country are generally more politically partisan or politically mixed. From: R.A. Arguedas, C. T. Robertson, R. Fletcher and R.K. Nielsen, Echo Chambers, Filter Bubbles, and Polarisation: a Literature Review, Reuters Institute for the Study of Journalism, 13 January 2022, accessed 13 January 2023.

[60] R.A. Arguedas, C. T. Robertson, R. Fletcher and R.K. Nielsen, Echo Chambers, Filter Bubbles, and Polarisation: a Literature Review, Reuters Institute for the Study of Journalism, 19 January 2022, accessed 13 January 2023.

[61] D. Garzia and F. Ferreria da Silva, Political polarisation means more Americans are voting against rather than for candidates in presidential elections, London School of Economics Blogs, 28 June 2022, accessed 17 January 2023.

[62] Centre for Data Ethics and Innovation, Online targeting: Final report and recommendations, February 2020, accessed 12 January 2023.

[63] ACMA, A report to government on the adequacy of digital platforms' disinformation and news quality measures, June 2021, accessed 13 January 2023.

[64]T. Lymn and J. Bancroft, The use of algorithms in the content moderation process, Centre for Data Ethics and Innovation Blog, 5 August 2021, accessed 12 January 2023.

[65] R. Gorwa, R. Binns, and C. Katzenbach, Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7:1, 2020.

[66] For example, several users on Twitter and Reddit whose accounts were suspended simply because they happened to share some of the key characteristics of disinformation purveyors. One user had tried to notify Twitter of a pro-Kremlin campaign, but ended up being banned himself. See: Marcechal, N & E.R. Biddle, It's Not Just the Content, It's the Business Model: Democracy's Online Speech Challenge, New America, 17 March 2020, accessed 13 January 2023.

[67] R. Gorwa, R. Binns, and C. Katzenbach, Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7:1, 2020.

[68] K. Conger, Jack Dorsey says Twitter played a role in U.S. Capitol riot, The New York Times, 25 March 2021, accessed 12 April 2023.

[69] R. Mac, C. Silverman and J. Lytvynenko, Facebook Stopped Employees From Reading An Internal Report About Its Role In The Insurrection. You Can Read It Here, BuzzFeed News, 26 April 2021, accessed 12 April 2023.

[70] E. Victoria and C. Stokel-Walker, Twitter's moderation system is in tatters, Wired, 17 November 2022, accessed 12 January 2023.

[71] Free Press, Empty promises: inside big tech's weak effort to fight hate and lies in 2022, 27 October 2022, accessed 12 January 2023.

[72] Ofcom, Media Plurality and Online news, November 2022, accessed 12 January 2023.

[73] G. Eady, T. Paskhalis, J. Zilinsky, R. Bonneau, J. Nagler & J.A. Tucker, Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior, Nature Communications, 14:62 (2023).

[74] eSafety Commissioner, Women in the Spotlight: Women's experiences with online abuse in their working lives. Melbourne, 2022, accessed 15 May 2023.

[75] T. Gillespie, Content moderation, AI, and the question of scale, Big Data & Society, 7:2 (2020)

[76] R. Griffin, Rethinking Rights in Social Media Governance, Verfassungsblog, 22 Feb 2022, accessed 12 January 2023.

[77] eSafety Commissioner, Position statement - Recommender systems and algorithms, December 2022, Centre for Data Ethics and Innovation, Online targeting: Final report and recommendations, February 2020, accessed 12 January 2023. Other similar definitions have also been proposed. For example, Park et al (2012) describes recommender systems as 'programs or systems designed to suggest [to a] user the next activity to indulge in, based on their preferences, history or a variety of other factors' D.H. Park, H.K. Kim, I. Y. Choi, J.K. Kim, A literature review and classification of recommender systems research, *Expert systems with applications,* 39:11 (2012).

[78] For example, a platform focused on rapid growth may focus on short-term engagement metrics like clicks, whereas a platform focused on maximising long-term use (leading to a greater number of overall clicks) may adopt metrics that could correlate with a higher quality experience, such as the length of time a user spends reading an individual article – Centre for Data Ethics and Innovation, Online targeting: Final report and recommendations, February 2020, accessed 12 January 2023.

[79] Centre for Data Ethics and Innovation, Online targeting: Final report and recommendations, February 2020, accessed 12 January 2023.

[80] Centre for Data Ethics and Innovation, Online targeting: Final report and recommendations, February 2020, accessed 12 January 2023.

[81] As well as with limited human review and editorial oversight – eSafety Commissioner, Position statement - Recommender systems and algorithms, December 2022.

[82] M. Rajibul, H. Ashish, K. Jha & Y. Liu, Excessive use of online video streaming services: Impact of recommender system use, psychological factors, and motives, *Computers in Human Behavior,* 80, March (2018), p. 220-228.

[83] J. Cobbe and J. Singh, Regulating Recommending: Motivations, Considerations, and Principles, *European Journal of Law and Technology*, 10:3 (2019).

[84] J. Cobbe and J. Singh, Regulating Recommending: Motivations, Considerations, and Principles, *European Journal of Law and Technology*, 10:3 (2019).

[85] J. Cobbe and J. Singh, Regulating Recommending: Motivations, Considerations, and Principles, *European Journal of Law and Technology*, 10:3 (2019).

[86] Centre for Data Ethics and Innovation, Online targeting: Final report and recommendations, February 2020, accessed 12 January 2023.

[87] Content-based filtering systems recommend content based on its similarity to content previously consumed by the user ("picture X has a similar title to previously viewed pictures Y and Z"), while Collaborative filtering systems recommend content based on what similar users have consumed ("people A, B and C like this; a similar person D might also like this"). Some platforms use hybrid approaches combining both methods.

[88] Ofcom, Media Plurality and Online news , November 2022, accessed 11 January 2023; CMA, Online platforms and digital advertising market study, Appendix F: the role of data in digital advertising, 1 July 2020, p. F19, accessed 11 January 2023.

[89] eSafety Commissioner, Position statement - Recommender systems and algorithms, December 2022.

[90] R. Slaughter, J. Kopac and M. Batal, Algorithms and Economic Justice: A Taxonomy of Harms and a Path Forward for the Federal Trade Commission, ISP Digital Future Whitepaper and Yale Journal of Law & Technology Special Bulletin, August 2021.

[91] R. Slaughter, J. Kopac and M. Batal, Algorithms and Economic Justice: A Taxonomy of Harms and a Path Forward for the Federal Trade Commission, ISP Digital Future Whitepaper and Yale Journal of Law & Technology Special Bulletin, August 2021.

[92] Australian Human Rights Commission, Using AI to make decisions: Addressing the problem of algorithmic bias, 2020.

[93] Australian Human Rights Commission, Human rights and technology – final report, 2021.

[94] Centre for Data Ethics and Innovation, Review into bias in algorithmic decision making, November 2020, accessed 12 January 2023. For example, see the discussion of proxy discrimination in R. Slaughter, J. Kopac and M. Batal, Algorithms and Economic Justice: A Taxonomy of Harms and a Path Forward for the Federal Trade Commission, ISP Digital Future Whitepaper and Yale Journal of Law & Technology Special Bulletin, August 2021.

[95] eSafety Commissioner, Position statement - Recommender systems and algorithms, December 2022.

[96] For example, studies by Allcott et al. (2020), Levy (2021) and Ofcom's own research suggest a link between social media and polarisation.

H. Allcott, L. Braghieri, S. Eichmeyer, and M. Gentzkow, The welfare effects of social media, American Economic Review, 110:3 (2020), p.629-676; R. Levy, Social Media, News Consumption, and Polarization: Evidence from a Field Experiment, American Economic Review, 111:3 (2021), p. 831-870; Ofcom, Media Plurality and Online news, November 2022, accessed 11 January 2023.

On the other hand, Nechushtai and Lewis (2019) examined whether news recommendation engines contribute to filter bubbles by asking a diverse set of users to search Google News for news about Hillary Clinton and Donald Trump during the 2016 U.S. presidential campaign and report the first five stories they were recommended on each candidate. Users with different political leanings from different states were recommended very similar news, challenging the assumption that algorithms necessarily encourage echo chambers. E. Nechushtai and S. C.Lewis What kind of news gatekeepers do we want machines to be? Filter bubbles, fragmentation, and the normative dimensions of algorithmic recommendations, *Computers in Human Behaviour*, 90 (2019), p. 298–307.

[97] Centre for Data Ethics and Innovation, Online targeting: Final report and recommendations, February 2020, accessed 12 January 2023.

[98] J. Reis, F. Benevenuto, P. Olmo, R. Prates, H. Kwak & J. An, Breaking the News: First Impressions Matter on Online News, in 'Proceedings of the Ninth International AAAI Conference on Web and Social Media, 357, 2015, pp 357-366.

S. Bradshaw & P. Howard, Why Does Junk News Spread So Quickly Across Social Media?, Oxford Internet Institute, 23 March 2018, accessed 13 January 2023.

[99] P. Dizike, Study: On Twitter, false news travels faster than true stories, MIT news, 8 March 2018, accessed 13 January 2023.

[100] The Center for Countering Digital Hate, Toxic Twitter, 9 February 2023, accessed 13 January 2023.

[101] J. Cobbe and J. Singh, Regulating Recommending: Motivations, Considerations, and Principles, *European Journal of Law and Technology*, 10:3 (2019).

[102] Z. Tufekci, Algorithmic Harms Beyond Facebook and Google: Emergent Challenges of Computational Agency, Colorado Technology Law Journal, 13 (2015), p. 203-217; J. Cobbe and J. Singh, Regulating Recommending: Motivations, Considerations, and Principles, *European Journal of Law and Technology*, 10:3 (2019).

[103] Ofcom, Media Plurality and Online news, November 2022, accessed 11 January 2023.

[104] Centre for Data Ethics and Innovation, Online targeting: Final report and recommendations, February 2020, accessed 12 January 2023.

[105] L. Molyneux and M. Coddington, Aggregation, Clickbait and Their Effect on Perceptions of Journalistic Credibility and Quality, *Journalism Practice*, 14:4 (2020), p.429-446.

[106] J. Cobbe and J. Singh, Regulating Recommending: Motivations, Considerations, and Principles, *European Journal of Law and Technology*, 10:3 (2019).

[107] Centre for Data Ethics and Innovation, Online targeting: Final report and recommendations, February 2020, accessed 12 January 2023.

[108] R. Fredheim and S. Bay, How Social Media Companies are Failing to Combat Inauthentic Behaviour Online, NATO Strategic Communications Centre of Excellence, 5 December 2019, accessed 12 January 2023.

[109] S. Bradshaw, H. Bailey and P.N. Howard, Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation, Oxford Internet Institute, February 2021, accessed 12 January 2023.

[110] Royal Commission of Inquiry into the Terrorist Attack on Christchurch Mosques on 15 March 2019, December 2020, accessed 4 February 2023.

[111] Russell Brandom, Inside Elsagate, the conspiracy-fueled war on creepy YouTube kids videos, The Verge, 8 December 2017, accessed 4 February 2023.

[112] Reset Australia, Designing for Disorder: Algorithms amplify pro-anorexia content to teens and children as young at 10, 22 April 2022, accessed 4 February 2023.

[113] eSafety Commissioner, Position statement - Recommender systems and algorithms, December 2022.

[114] Y. Gerrard & T. Gillespie, When Algorithms Think You Want To Die, Wired, 21 February 2019, accessed 19 January 2023.

[115] For example, the suicide of Molly Russell in the UK in 2017 highlighted concerns about the link between recommendations of self-harm content and negative mental health, as she reportedly binge consumed self-harm related content served by recommender systems.

A recent scoping review of relevant literature by Moss et al (2022) indicated a relationship between time spent on Instagram and deliberate self-harm; desensitization of deliberate self-harm resulting in normalization; social contagion and that Instagram provided a sense of belonging to its users who engaged in deliberate self-harm. See C. Moss, C. Wibberley and G. Witham, Assessing the impact of Instagram use and deliberate self-harm in adolescents: A scoping review, *International Journal of Mental Health Nursing*, 32:1 (2022), p.14-29.

[116] eSafety Commissioner, Position statement - Recommender systems and algorithms, December 2022.

[117] Digital Regulators Co-operation Forum, The benefits and harms of algorithms: a shared perspective from the four digital regulators, 23 September 2022, accessed 4 February 2023.

[118] M. Rajibul, H. Ashish, K. Jha & Y. Liu, Excessive use of online video streaming services: Impact of recommender system use, psychological factors, and motives, *Computers in Human Behavior,* 80, March (2018), p. 220-228.

[119] ACCC, Trivago to pay $44.7 million in penalties for misleading consumers over hotel room rates, 22 April 2022, accessed 12 January 2023.

[120] eSafety Commissioner, Position statement - Recommender systems and algorithms, December 2022

[121] J Cobbe and J Singh, Regulating Recommending: Motivations, Considerations, and Principles, *European Journal of Law and Technology*, 10(3), 2019.

[122] New York Times, The making of a YouTube radical, 8 June 2019, accessed 12 January 2023.

[123] eSafety Commissioner, Position statement - Recommender systems and algorithms, December 2022.

[124] OAIC, 2020 Australian Community Attitudes to Privacy Survey, September 2020, p 29.

[125] OAIC, 2020 Australian Community Attitudes to Privacy Survey, September 2020, p 31.

[126] A. Descamps, T. Klein and G. Shier, Algorithms and competition: the latest theory and evidence, *Competition Law Journal*, 20:1 (2022).

[127] General Court of the European Union, The General Court largely dismisses Google's action against the decision of the Commission finding that Google abused its dominant position by favouring its own comparison shopping service over competing comparison shopping services, Press release, 10 November 2021.

The ACCC has similarly raised concerns about the potential for anti-competitive self-preferencing in the context of digital platforms. ACCC, Digital Platform Services Inquiry – Interim report no. 5 – regulatory reform, November 2022, See Chapter 6.1.

[128] ACCC, Digital advertising services inquiry – Final report, August 2021.

[129] Centre for Data Ethics and Innovation, Online targeting: final report and recommendations, 4 February 2020, accessed 12 January 2023.

[130] A. Nadler, M. Crain and J. Donovan, Weaponizing the digital influence machine: the political perils of online ad tech, *Data & Society*, 2018, accessed 17 January 2023.

[131] A. Nadler, M. Crain and J. Donovan, Weaponizing the digital influence machine: the political perils of online ad tech, *Data & Society*, 2018, accessed 17 January 2023; ACCC, Digital advertising services inquiry – Final report, August 2021; Competition and Markets Authority (CMA), Online platforms and digital advertising market study final report, 1 July 2020; Salinger Privacy and Castellate Consulting, Cookies and other online identifiers: research paper for the Office of the Australian Information Commissioner, June 2020, accessed 16 January 2023; M Ali et al, Discrimination through optimization: how Facebook's ad delivery can lead to biased outcomes, *Proceedings of the ACM on Human-Computer Interaction*, 2019, 3(CSCW):1–30, accessed 17 January 2023; P Sapiezynski et al, Algorithms That "Don't See Color": Comparing Biases in Lookalike and Special Ad Audiences*, Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society,* 2022, accessed 20 January 2023.

[132] A. Nadler, M. Crain and J. Donovan, Weaponizing the digital influence machine: the political perils of online ad tech, *Data & Society*, 2018, accessed 17 January 2023; ACCC, Digital advertising services inquiry – Final report, August 2021; CMA, Online platforms and digital advertising market study final report, 1 July 2020, Appendix G; Salinger Privacy and Castellate Consulting, Cookies and other online identifiers: research paper for the Office of the Australian Information Commissioner, June 2020, accessed 16 January 2023; J.M. Paterson, S. Dreyfus and S. Chang, What we see and what we don't: protecting choice for online consumers policy report, The University of Melbourne, 2020, accessed 20 January 2023.

[133] M. Andrejevic et al, Unregulated and Segmented Dark Ads on Social Media Consumer Education and Regulatory Options*,* Monash Automated Society Working Group, 2020, accessed 20 January 2023; P. Sapiezynski et al, Algorithms That "Don't See Color": Comparing Biases in Lookalike and Special Ad Audiences*, Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society,* 2022, accessed 20 January 2023*;* A. Kofman and A. Tobin, Facebook Ads Can Still Discriminate Against Women and Older Workers, Despite a Civil Rights Settlement, *ProPublica,* 13 December 2019, accessed 20 January 2023; M. Ali et al*, Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes*, *Proceedings of the ACM on Human-Computer. Interaction*, 2019, 3(CSCW):1–30, accessed 17 January 2023. R. Slaughter, J. Kopac and M. Batal,

Algorithms and Economic Justice: A Taxonomy of Harms and a Path Forward for the Federal Trade Commission, ISP Digital Future Whitepaper and Yale Journal of Law & Technology Special Bulletin, August 2021.

[134] Ibid.

[135] Salinger Privacy and Castellate Consulting, Cookies and other online identifiers: research paper for the Office of the Australian Information Commissioner, June 2020, accessed 16 January 2023.

[136] M. Andrejevic at al, Unregulated and Segmented Dark Ads on Social Media Consumer Education and Regulatory Options, Monash Automated Society Working Group, 2020, accessed 20 January 2023; J. Angwin, M. Varner and A. Tobin, Facebook Enabled Advertisers to Reach 'Jew Haters', *Propublica*, 14 September 2017, accessed 3 January 2023.

[137] A. Kofman and A. Tobin, Facebook Ads Can Still Discriminate Against Women and Older Workers, Despite a Civil Rights Settlement, *ProPublica,* 13 December 2019, accessed 20 January 2023.

[138] M. Ali et al, Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes, *Proceedings of the ACM on Human-Computer Interaction*, 2019, 3(CSCW):1–30, accessed 17 January 2023; Centre for Data Ethics and Innovation, Online targeting: final report and recommendations, 4 February 2020, accessed 12 January 2023.

[139] A. Lambrecht and C. Tucker, Algorithm-Based Advertising: Unintended Effects and the Tricky Business of Mitigating Adverse Outcomes, 2021, accessed 20 January 2023; M. Ali et al, Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes, *Proceedings of the ACM on Human-Computer Interaction*, 2019, 3(CSCW):1–30, accessed 17 January 2023.

[140] M. Andrejevic et al, Unregulated and Segmented Dark Ads on Social Media: Consumer Education and Regulatory Options, Monash Automated Society Working Group, 2020, accessed 20 January 2023.

[141] Ibid.

[142] Information Commissioner's Office (UK), Update report into adtech and real time bidding, 20 June 2019, accessed 17 January 2023; ACCC, Digital advertising services inquiry – Final report, August 2021; CMA, Online platforms and digital advertising market study final report, 1 July 2020.

[143] ACCC, Digital advertising services inquiry – Final report, August 2021.

[144] J.M. Paterson et al, The Hidden Harms of Targeted Advertising by Algorithms and Interventions from the Consumer Protection Toolkit, *International Journal of Consumer Law and Practice*, 2021, 9:1-17, accessed 20 January 2023. Advertising Standards Authority, Harnessing new technology to tackle irresponsible gambling ads targeted at children, 2019. A. Brownbill, Op-ed: How digital marketing of alcohol, gambling and junk food harms consumers, Choice, 27 May 2022.

[145] J. Burgess, M. Andrejevic, D. Angus and A.K. Obeid, Australian Ad Observatory: background paper, ADMS, 2022, accessed 20 January 2023.

[146] J.M. Paterson et al, The Hidden Harms of Targeted Advertising by Algorithms and Interventions from the Consumer Protection Toolkit, *International Journal of Consumer Law and Practice*, 2021, 9:1-17; accessed 20 January 2023.

[147] DQube Solutions et al., Drawing back the curtain: consumer choice online in a data tracking world, 2020, accessed 17 January 2023.

[148] Ibid; M. Andrejevic, D. Angus and J. Burgess, Why we need better oversight of targeted online advertising, *Choice*, 8 April 2022, accessed by 17 January 2023.

[149] K. Hawker et al., Advertisements on digital platforms: How transparent and observable are they?, *Foundation for Alcohol Research & Education*, 2 September 2022, accessed 17 January 2023; J. Burgess, M. Andrejevic, D. Angus, A.K. Obeid, Australian Ad Observatory: background paper, ADMS, 2022, accessed 20 January 2023; M. Ali et al, Discrimination through optimization: how Facebook's ad delivery can lead to biased outcomes, *Proceedings of the ACM on Human-Computer Interaction*, 2019, 3(CSCW):1–30, accessed 17 January 2023; B. Carlson and M. Andrejevic, There is a long

history of racist and predatory advertising in Australia. This is why targeted ads could be a problem, *The Conversation*, 18 October 2021, accessed 17 January 2023.

[150] Ibid; M. Andrejevic, D. Angus and J. Burgess, 'Why we need better oversight of targeted online advertising', *Choice*, 8 April 2022, accessed 17 January 2023

[151] See for example W. Dunn, Anti-vaccination advert banned - but Facebook still offers targeting of people susceptible to 'vaccine controversies', *New Statesman*, 7 November 2018, accessed 23 January 2023; A. Nadler, M. Crain and J. Donovan, Weaponizing the digital influence machine: the political perils of online ad tech, *Data & Society*, 2018, accessed 17 January 2023.

[152] T. Tilley, Tom Tilley interviews Professor Daniel Angus: Election ads: from Google to Grindr [interview audio file], The Briefing, May 2022, accessed 17 January 2023.

[153] A. Nadler, M. Crain and J. Donovan, Weaponizing the digital influence machine: the political perils of online ad tech, *Data & Society*, 2018, accessed 17 January 2023; Jeremy B Merrill, What we learned from collecting 100,000 targeted Facebook ads, *Propublica*, 26 December 2018, accessed 17 January 2023.

[154] T. Tilley, Tom Tilley interviews Professor Daniel Angus: Election ads: from Google to Grindr [interview audio file], *The Briefing*, May 2022, accessed 17 January 2023; A. Nadler, M. Crain and J. Donovan, Weaponizing the digital influence machine: the political perils of online ad tech, *Data & Society*, 2018, accessed 17 January 2023.

[155] A. Nadler, M. Crain and J. Donovan, Weaponizing the digital influence machine: the political perils of online ad tech, *Data & Society*, 2018, accessed 17 January 2023.

[156] C. Newton, How Facebook rewards polarising political ads, *The Verge*, 12 October 2017, accessed 17 January 2023.

[157] A. Nadler, M. Crain and J. Donovan, Weaponizing the digital influence machine: the political perils of online ad tech, *Data & Society*, 2018, accessed 17 January 2023.

[158] Wired, Meet the lobbyist next door,17 July 2022, accessed 17 January 2023.

[159] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)

[160] European Commission, 2022 Strengthened Code of Practice on Disinformation, 16 June 2022

[161] Federal Register of Legislation, Online Safety (Basic Online Safety Expectations) Determination 2022, 23 January 2022.

[162] Medium, China has pioneered a law to empower people over algorithms, 9 March 2022, accessed 31 January 2023.

[163] Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act)

[164] Department of Industry, Science and Resource, Australia's AI Ethics Principles, 7 November 2019

[165] OECD, OECD AI Principles, May 2019.

[166] European Commission, Ethics guidelines for trustworthy AI, 8 April 2019

[167] UNESCO, Recommendation on the Ethics of Artificial Intelligence, 2022.