Position statement

Recommender systems and algorithms

Australian Government

eSafetyCommissioner

esafety.gov.au

Contents

December 2022

Overview

The eSafety Commissioner (eSafety) is Australia's independent regulator and educator for online safety – the first of its kind in the world. eSafety represents the Australian Government's commitment to protecting citizens from serious online harms.

To ensure eSafety's content and programs reflect current information, technological developments and global trends, eSafety scans for new research, policy, and legislative and technical updates. We have captured eSafety's approach to selected tech trends and challenges in our position statements. This position statement is focused on recommender systems.

Definition

Recommender systems, also known as content curation systems, are the systems that prioritise content or make personalised content suggestions to users of online services.

A key component of a system is its recommender algorithm, the set of computing instructions that determines what a user will be served based on many factors. This is done by applying machine learning techniques to the data held by online services, to identify user attributes and patterns and make recommendations to achieve particular goals.

Background

Recommender systems and their underlying algorithms are integral to many online services. These systems sort through vast amounts of data to present content that is relevant, interesting, and suitable to users. This helps people discover new artists, friends, products, activities, and ideas as well as helping businesses and creators reach new audiences.

For these reasons, recommender algorithms are used by online services such as Facebook, Instagram, TikTok, Twitter and YouTube to amplify, prioritise and recommend content and accounts to their users. Recommender algorithms are also used to deliver relevant search results on many different types of online services, including search engines.

The data used by recommender algorithms to produce suggestions for users can be sourced in several ways. It may be provided knowingly by users through actions such as search queries, ranking posts or providing feedback on their preferences. Alternatively, it may be captured by collecting information about a user, such as their demographic details, or by monitoring their engagement with content, such as likes, comments or dwell time (how long users hover over specific content before scrolling past). In some cases, it may be purchased from other third-party data sellers or services.

Drawing on this data, recommender algorithms can be optimised to achieve different purposes. For example, a service may be seeking to maximise user engagement through maximising time spent on its service, 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.

Different inputs and end goals for recommender systems can lead to both positive and negative outcomes. For example, recommender algorithms that prioritise time spent reading or reacting to a post and then serve up similar content in the future may result in people seeing things they find interesting, entertaining or valuable. But equally, if a user spends time engaging with potentially harmful content, those same metrics may lead to individuals seeing more of the same material or increasingly harmful material in their feeds.

Notably, the question of whether content is harmful can depend on the individual user and their context and circumstances. For example, content that promotes self-harm is likely to present a greater risk to someone already experiencing mental ill health.

In addition to contributing to risks and harms at an individual level, recommender systems have the potential to cause or exacerbate harms on a societal level. For example, content promoting hate or inciting violence can cause damage to the people targeted and can also spill over into violence affecting the broader community.

The level of risk can also be impacted by factors specific to the online service, such as the size and quality of its pool of content, as well as external factors, such as broader socio-political developments.

For these reasons, it is important to take a holistic view of recommender systems which encompasses their benefits and risks, broader uses, and complex interconnected ecosystems.

Online safety risks and considerations

While recommender systems are used across many technologies to great benefit, they also present an array of risks to online safety, particularly in the context of social media and other online services that enable user generated content to be shared.

One of these risks is the potential for recommender systems to **amplify harmful and extreme content**.

On an individual level, this can increase the impact on those exposed to harmful material.

On a broader societal level, the amplification of content that promotes discrimination, such as sexism, misogyny, homophobia or racism, can have adverse effects by normalising prejudice or hate. This may also contribute to **radicalisation** towards terrorism and violent extremism.

The Centre for Countering Digital Hate report titled The Incelosphere outlines how mainstream services act as pathways to incel (involuntarily celibate) communities, for example, through YouTube videos and Google search results. This research shows users seeking body image content online is a key pathway to incel communities.

One key driver of this risk comes from the way social media services **optimise** their recommender systems for greater **engagement**.

Systems designed to recommend content with the purpose of increasing engagement run the risk of exploiting human cognitive biases, inherently drawing people to shocking and extreme content. This may lead to services recommending content that comes close to violating, or is definitively in breach of, their terms of service, community standards or local law.

As well as potentially amplifying harmful content or inappropriately targeting users, services' recommender systems are routinely **exercising content moderation** decisions, whether in order to throttle the reach of 'borderline content'[1] for all users, or to direct particular kinds of content away from users who might be offended by it. The extent to which these decisions are explainable, observable, measurable, or reported is very limited and raises transparency issues.

'Shadowbanning' (or surreptitiously suppressing) borderline content or accounts that share this content can help slow its spread. However, this can be problematic where it occurs suddenly, without warning or explanation, and with limited options for review or recourse. Shadowbanning can disproportionately impact content creators who deal with sensitive subject matter, such as content relating to sex, self harm, drugs or violence, even where the purpose is to provide education or information.

Business models can also affect risk. Online services operating with **ad-based revenue models** are incentivised to increase engagement as the more time users spend online, the more advertising revenue is generated.

While traditional media outlets have been drawing on inflammatory content to increase their audiences for many years before recommender algorithms became ubiquitous online, the risk of harm may be more pronounced online.

Dominant digital services present a unique challenge in that their recommender systems can amplify content to global user bases with **limited human review and editorial oversight**. In addition, the enormous and ever-expanding inventory of user-generated content on large global services means they have more harmful content available to disseminate.

Increasingly, traditional and social media also feed off one another, with social media being used to promote stories from traditional media outlets, and those outlets reporting on activities occurring on social media.

Humans who use these services are also key actors in the posting and sharing of potentially harmful content. Their input and behaviour, in conjunction with other elements such as how a user interface is designed, what content is posted by media publications, and broader cultural norms, all play an important part in how content is recommended.

Extreme content which people may choose to share and engage with online is often rooted in **broader societal issues** or polarising and emotionally charged debates, such as on political issues, rather than simply a service's algorithm. In addition, people deliberately seeking out online communities with extreme views have avenues to find these groups irrespective of the type of content that is recommended to them.

However, recommender algorithms are significant in that they can give disproportionate reach to harmful content posted by a minority of users. Some users may even feel incentivised to exploit a recommender system by making their material more inflammatory or extreme.

Amplification of any content necessarily involves giving it preference over other content. By amplifying some content, recommender systems can end up deprioritising or excluding different viewpoints or valuable ideas contrary to a person's existing beliefs, contributing to what are commonly known as **echo chambers** or **filter bubbles**. Echo chambers can impact a person's freedom of thought, access to information and autonomy, and can contribute to polarisation.

While there is some evidence linking recommender systems to filter bubbles, experts have also highlighted that these issues may be overstated.[2]

In situations of livestreamed violent extremism, such as the 2019 terrorist attack on two Christchurch mosques, perpetrators can access global audiences, causing widespread harm to unsuspecting users. The 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.

Certain individuals or groups may be more likely to be served potentially harmful material, due to their demographics and/or their previous online activity. The severity and type of harm they may experience because of those recommendations will vary depending on their circumstances.

These risks and harms can be exacerbated by **intersectional factors**. A piece of content can be harmful for some people and communities and not others. This can make it challenging to prevent specific instances of harm or measure the severity of harms caused by recommender systems.

eSafety's research shows that some people are at greater risk of experiencing online abuse than others. Recommender systems can accentuate this risk by increasing the virality of harmful content and laying the foundation for **pile-ons or volumetric attacks** to take place.

Children and young people in particular face risks from recommender systems, including the following:[3]

- Exposing them to heightened risks of online sexual exploitation through friend/follower suggestions that pressure children to interact with potentially dangerous adults. There have also been reports of recommender systems serving sexualised videos of children to perpetrators, and enabling them to exchange information with other predators through comment sections. Some take this a step further by adding a child's social media account to groom them into producing more sexualised content.[4]
- Drawing them into content loops such as those on video-sharing services, which can limit their exposure to diverse content, deliver increasingly problematic content or result in extended periods of time online. Recent media coverage has highlighted how social media recommender systems can contribute to negative mental health impacts. The coroner's report in the inquest into Molly Russell's death found that social media was a causal factor in her suicide. This stemmed in part from how Molly binge consumed self-harm related content through algorithmically curated feeds.[5]
- Encouraging dangerous viral challenges, such as the 'blackout challenge' seen on TikTok which has led to multiple deaths of children[6], or viral memes that constitute cyberbullying.
- Promoting 'ideals' of body types and beauty stereotypes or normalising the sexualisation of young people. This includes previous reports of sexualised content being recommended to children.[7]
- Recommending content that may be appropriate for adults but harmful to children who are not developmentally ready for it, such as violent or sexually explicit material.

Other risks and regulatory concerns

In addition to the online harms that fall within eSafety's regulatory remit, recommender systems raise several cross-cutting issues of concern to various government departments, agencies and regulators, both within Australia and worldwide.

While primary responsibility for many of these issues sits with other organisations, they have the potential to intersect with the online harms within eSafety's remit.

Relevant cross-cutting issues related to recommender systems include the following:

Competition and consumer issues

For example, using information or data to influence competition or consumer choice or conduct.

- The use of recommender algorithms brings a range of benefits and can support competition and consumer outcomes, for example, by **reducing search costs** for consumers when purchasing products online.
- However, these algorithms can also pose risks or issues for competition and consumers. For example, recommender systems **can encourage the sharing or the spread of scam content** or **mis or disinformation**.
- Moreover, recommender algorithms may be used by firms to engage in exclusionary conduct (restricting or undermining their rivals' ability to compete in order to maintain or advance their own position in the market). For example, issues of self-preferencing could arise on online marketplaces or app stores that include a service's own offerings alongside competing third-party offerings in search results. Given that the order and prominence with which products are displayed can have a substantial influence on which products consumers select, a service could benefit by using recommender algorithms that give priority to their own products at the expense of third-party competitors. This preferential treatment could inhibit the ability of third parties to compete and consumers may suffer if they expect the highest ranked products to be those that best meet their search criteria or needs.

Communications and media issues

For example, influencing the modern news and information environment.

- **Access to news and other content**. Recommender algorithms can broaden consumer engagement and choice by providing diversity and access to news content. However, to drive engagement and attention, users are often driven towards content they are likely to engage based on previous engagement, user preferences or interests. Recommender systems paired with self-selection may drive users into information filter bubbles and echo chambers. These siloing forms of engagement may be harmful to users and to society, by removing counter-attitudinal/authoritative information, localised content, and media diversity in favour of perceived personalisation.
- **Misinformation**. Recommender systems have the potential to promote misinformation, 'borderline' or 'fringe' content that may not be factually accurate or scientifically validated. Such content might create confusion and reduce trust, but may not be directly in breach of a service's terms of service, and therefore may not be removed.
- **Polarisation or political interference**. Recommender systems could influence users' political views and democratic outcomes by promoting certain political content over other information.
- **Harmful advertising and targeted advertising**. This might include harmful advertising relating to online gambling, tobacco or alcohol, as well as scams. Harmful advertisements can target high risk communities based on a range of signals such as behavioural patterns.

Privacy and discrimination issues

For example, using personal information or data to influence recommendations or choice.

- **Privacy**. The collection and use of personal information, including biometrics, to produce recommendations may impact privacy. For example, personal information might be collected to build a profile that can be used in targeted advertising or direct marketing to influence consumer choice. Although more detailed profiles may improve the accuracy of the recommendations, this can lead to greater collection of personal information and increased privacy risks. To build this profile, recommender systems may infer additional information about individuals, such as their interests. Use of personal information in recommender systems should comply with privacy obligations and individuals should be made aware of how their personal information is being collected and used by the recommender system.
- **Bias and discrimination**. Some individuals and communities experience harms from recommender systems at disproportionate rates. This stems in part from how recommender systems are powered by user data reflecting human biases, and how they can amplify those biases. This can cause recommender systems to discriminate against diverse communities when content representing these groups is excluded from recommendations, or even reinforce and amplify harmful stereotypes, biases, or beliefs through targeted advertising and content recommendation. For example, a recommender system might prioritise content with what they determine as a higher 'beauty ranking', failing to reflect ethnically diverse users and reinforcing homogenous beauty standards. It can also mean that content creators from minority groups are less likely to have their content recommended. Amplification of biases may also lead to an increased likelihood of online abuse, hate speech and 'pile-ons' or 'volumetric attacks'.

Regulatory challenges

Any regulation of recommender systems must safeguard the rights of users while preserving the benefits of these systems and fostering healthy innovation.

Recommender algorithms are in constant flux; they change over time responding to user input, economic opportunities, regulation and public sentiment. Primarily, effective regulatory responses require greater knowledge and understanding of how these systems operate. Without this information, regulatory interventions may generate adverse consequences. At the same time, accessing information requires both a legislative basis and careful consideration for business impacts.

Transparency

Online services' use of algorithms are typically opaque. Greater transparency is critical to improving both regulatory and public understandings of algorithms and to making sure online services are accountable for the impact of their design choices. Transparency by their creators in relation to what algorithms and recommender systems can do, and are intended to do, is the most sustainable approach to enabling regulatory oversight and empowering people to make informed choices.

There are significant challenges in explaining the functionality of complex algorithmic decision-making systems and their rationale in specific cases, as discussed in the Ada Lovelace Institute's paper from December 2021 on technical methods for regulatory inspection of algorithmic systems.

One common issue is that explanations can be difficult to understand, as can the code that algorithms are comprised of. Recommender algorithms can be understood as an element within a broader content moderation ecosystem. However, this makes transparency and accountability efforts difficult because these measures are harder to report and observe than content removal actions.

Globally, regulatory efforts seeking to address these issues are emerging. This includes eSafety's transparency reporting powers under the Basic Online Safety Expectations to make sure online safety measures are considered and employed, as well as the EU's Digital Services Act which has data access obligations for very large services. Regulators are in a unique position to establish longer-term infrastructure that can support research and investigations at a broader level over time.

There is also potential for regulators and services across jurisdictions to collaborate with academics and researchers and to provide access to information. This could improve public understandings of algorithms, enhance online safety and minimise harms.

Evolving technologies

The rapidly evolving nature of technology presents another challenge to regulation.

Emerging technologies will also converge with each other. In doing so, they have the potential to mutually reinforce safety risks. For example, recommender systems in an immersive environment or metaverse could risk exposing people to harmful content or targeted advertising in a much more visceral way.

Any future regulation that looks at developing technologies like recommender algorithms should take a technology-agnostic approach to combat a variety of online harms. This is consistent with global uptake of eSafety's Safety by Design initiative, which applies across a broad and comprehensive set of technologies, processes and services.

Other considerations

Important considerations for regulatory efforts targeted at recommender systems include:

- **Harmonising efforts** across global government agencies to avoid a fragmented regulatory environment and unnecessary duplication.
- Understanding the **underlying ad-based revenue models** which many large online services employ and aligning incentives so that safety considerations are considered in tandem with business incentives.
- **Enhancing education** around the existence, role and impact of algorithms on our interaction with news and other media, recognising that regulation alone cannot mitigate all the risks.

eSafety's approach

Prevention

Recognising the importance of enhancing digital literacy and giving people the skills and confidence to manage their online experiences, eSafety is developing education and training programs to help raise awareness of the potential risks of recommender systems and the tools to manage them.

Critical digital literacy is important in knowing algorithms exist and how they work, being able to identify instances of when users are being targeted, as well as understanding how an algorithm may be shaping an individual's experience, feelings or reasoning, and equips them with skills to have a role in their choices around news and other media.

eSafety encourages the development of critical thinking early, with scaffolded messaging and learning starting with our Early Years program for under 5s (and their parents and educators), through to young people (13 to 17), and beyond. eSafety's current approach to critical digital literacy includes incorporating relevant guidance into prevention primary and secondary programs. This includes being embedded in pre-service and teacher professional learning, webinars for parents and carers and outreach through our Trusted eSafety Provider program. We are also working closely with mental health professionals, child protection services, and other frontline workers to develop key resources for groups at risk of harm.

eSafety's 2022 Mind the Gap study showed that almost two-thirds of young people aged 14–17 were exposed in the preceding 12 months to potentially negative content or harmful content, such as content relating to drug taking, suicide or self-harm, or gory or violent material which may have been recommended to them. Additionally, one in ten children have been the target of hate speech online, and many parents underestimate the prevalence of children's negative online experiences. We engage with young people through an Online Safety Youth Advisory Council, where key issues like recommender systems and algorithms will be explored during the 24-month term of the Council. The Council will report to the Australian government on the challenges they face online and their aspirations and solutions for a safer and more positive online world.

eSafety's research shows that diverse communities experience online harms at disproportionate rates. For example, our Protecting voices at risk online report indicated that First Nations people and the LGBTIQ+ community experience online hate at twice the rate of the general population.[8] It is important to consider whether and how recommender systems may be fuelling these statistics by amplifying harmful content.

We are also promoting algorithmic literacy and understanding about how algorithms function and content is personalised, to help give people greater influence over their recommendations.[8] This includes special resources for groups at higher risk, especially children, for whom critical thinking skills and news and other media literacy are crucial. eSafety is also supporting research in this area. For example, eSafety is funding a University of Sydney-led project to co-design and produce social media safety videos for Instagram and TikTok tailored to young people.

Protection

eSafety takes steps to limit the availability and spread of harmful material through our investigations schemes. These schemes enable the Commissioner to take action to remove illegal and restricted content, cyberbullying content targeted at a child, serious adult cyber abuse and intimate images that have been shared without the consent of the person shown (known as 'image-based abuse').

eSafety uses its powers to stop the spread of harmful material from continuing to be amplified. In crisis situations, such as when the Buffalo shooting video and manifesto were circulated, eSafety was able to issue removal notices to websites hosting terrorist content to remove the content within 24 hours, and to request that search engines remove links to the content.

Proactive and systemic change

eSafety conducts consultation and horizon scanning to remain future focused, and ready for emerging issues and potential regulatory challenges and opportunities. This includes the systemic issues presented by the rapid development of recommender systems and algorithms. eSafety's research shows that some communities experience online harms at disproportionate rates. For example, our Women in the Spotlight report found that 6% of women were directly targeted by an 'anti' or 'hate' group whose membership and reach may be fuelled in part by recommender systems.

We guide and support industry to enhance online safety measures through our **Safety by Design** initiative, explained in further detail in the next section.

The *Online Safety Act 2021* enables eSafety to regulate the systems and processes of online services through the co-regulatory industry codes and the transparency and accountability powers relating to the Basic Online Safety Expectations.

Basic Online Safety Expectations

The Basic Online Safety Expectations give eSafety the power to require services to report on how they are meeting a range of expectations. This includes a safe use expectation, and an expectation to proactively minimise unlawful and harmful material and activity.

Although recommender systems are not specifically mentioned, eSafety considers that part of creating a safe environment for users is ensuring that recommender systems do not amplify unlawful and harmful material and activity. In order to understand how providers are implementing the Expectations, eSafety may ask questions relating to a recommender system's intended objectives and outcomes, and how services evaluate and adjust their algorithms to prevent harm to users.

Through ongoing research and consulting with experts, eSafety is also considering ways in which increased transparency and accountability measures, including through the Basic Online Safety Expectations, could incentivise companies to take proactive steps to ensure their recommender systems protect users.

Current industry practice

The use of recommender systems by digital services to determine search results and to suggest content for newsfeed style pages has become ubiquitous.

Recommendations are commonly derived through **content-based filtering** (where a user is matched to content based on having shown interest in the same category) or **collaborating filtering** (where a user is served content liked by other users classified as similar).

Recommender systems are enabled by machine learning and involve statistical processes and numerical proximations to reach an objective. These statistical processes are largely autonomous and adaptive and may fail to adequately assess the nature of content which they are recommending. The common practice of using signals such as likes, shares and dwell time as a measurement for engagement in lieu of more precise metrics can inadvertently amplify harmful content. In addition, there is typically limited transparency on how these signals are used and weighted in a recommender system.

However, services also optimise their recommender systems for metrics other than engagement, such as:

- **Diversity and explorability** – to recommend different types of content so users are exposed to different views
- **Authoritativeness** – to recommend content from credible sources.

Another common practise is to refrain from recommending 'borderline' or 'fringe' content, or material that comes close to violating terms of service but does not quite meet the threshold. This allows services to make content which could be harmful in aggregate less prominent instead of removing it completely.

However, there is a risk that automated systems which perform this downranking function may over-capture content that is not harmful. For example, a system calibrated to deprioritise sexually explicit content may inadvertently remove sexual health content from recommendations, making it difficult for users to find valuable information. This risk can also have a disproportionate effect on underrepresented or marginalised cultures and communities. In addition, interventions such as downranking are harder to measure and have less transparency than interventions such as content removals.

Guidance for industry

Users' safety, wellbeing, best interests and rights must be paramount considerations when developing and implementing recommender systems.

As is the case for any important component of an online service, recommender systems require active and consistent monitoring to detect and address risks as they arise. This includes through ongoing risk assessments, impact and outcome assessments, proactive content moderation measures, and transparency efforts. These measures should be applied across a product's full lifespan.

The limited **transparency** of how services use recommender systems impedes effective regulatory efforts and public awareness.

eSafety encourages services to make available information on how their recommender systems and algorithms operate. Rather than attempting to 'break the black box' – or crack open the service's algorithms to see what is happening on the inside – this should include information on design choices, an algorithm's objectives, and the positive and negative outcomes of recommender systems.

Researchers face various challenges when studying the effects of recommender systems, including barriers to having ongoing access to consistent data over time. Longitudinal studies are important because recommender systems are constantly changing and adapting. Researchers may also face legal challenges and so research protections to enhance research, evidence and transparency about online harms is important.[10]

While their constant state of flux presents challenges to both longitudinal studies and transparency, it means algorithmic audits and impact assessments must be carried out on an ongoing basis, rather than as one-off.

Sharing information and working collaboratively with researchers can be an effective means of detecting and evaluating risks and harms for users and broader society. It could also be a way of exploring how algorithms could be used to recommend content and services that build protective factors and facilitate help-seeking.

Further research from academics in local contexts would also benefit services' efforts, regulatory oversight and public awareness.

Proactive and automated tools that allow services to detect illegal and harmful content online are important for addressing harmful content at-scale and for preventing it from being spread through recommender systems. However, such **content moderation algorithms** also carry risks, such as unintentionally over-restricting speech.

Engagement can be a valuable metric for recommender systems. However, if too much emphasis is placed on engagement, or if the effects of engagement-based recommender algorithms are unchecked, this can risk harming people and communities.

Optimising recommender systems with other metrics can help to reduce the risk of amplifying harmful content. As services are responsible for designing and training their recommender algorithms, they are also responsible for designing systems which do not dangerously amplify harmful content. This can be done through adopting a Safety by Design approach and making sure the goals of the system align with the best of interests of users and creators.

eSafety's Safety by Design and proactive measures

eSafety encourages services to make available information on how their recommender systems and algorithms operate. Rather than attempting to 'break the black box' – or crack open the service's algorithms to see what is happening on the inside – this should include information on design choices, an algorithm's objectives, and the positive and negative outcomes of recommender systems.

How these measures are applied depends largely on the different considerations that apply to the context in which the recommender system is deployed, such as whether it is used in social media, search engines or another type of service. For instance, social media is designed for engagement while search engines are designed to deliver relevant results.

Interventions will also differ based on whether interventions are directed at individual harms or societal harms. For example, measures that give users greater control over their recommendations would be less impactful for dealing with societal issues.

A Safety by Design approach is critical to keeping users safe and building trust with communities. Services can consider a range of safety interventions to minimise the risk of harm from recommender systems, including:

- **Opt-out measures**. Recommender systems are a valuable way for people to discover new things, but they also exist as a business model to enhance engagement and profit. Users should have an option to maintain choice, ownership and control of the types of content they receive and be able to opt-out of others.
- **Adjusting recommender algorithms** to focus on more quality-focused metrics instead of or in addition to engagement, such as the authoritativeness or diversity of content. These metrics should be subject to consultation, public scrutiny and testing to avoid subjective processes in determining what sources are authoritative.
- **Offering users alternative curation models for their news feeds**, such as a reverse chronological news feed. It's also important that these options are designed in a way that is accessible and simple to use, and not built into complex interfaces. One approach receiving renewed attention is referred to as a type of 'middleware' solution. It would allow users to choose from a range of providers offering their own recommender and content moderation algorithms. In this context, middleware refers to third-party services offering editorial software that sits between users and services.[11] More research is needed on the adoption of these alternatives, and how effective they are in mitigating online safety risks.
- **Providing greater choice, control and clear feedback loops for users**. Services can empower users to explicitly shape their recommender systems, for example, by allowing them to flag types of content they do not want to see in suggested posts. It is important there is also transparency and consultation on the effectiveness of user tools and actions, as there is considerable scope for services to misinterpret the feedback users are trying to provide and the reasons why they want to opt out of particular types of content.
- **Establishing and enforcing content policies, and actively moderating harmful content** to make sure a platform's pool of available content meets a baseline threshold. Services with a greater density of harmful content available are more likely to amplify that harmful content.
- **Actively monitoring for and removing illegal content** that clearly violates relevant laws and community standards, such as child sexual exploitation material and terrorist content.

- **Introducing human reviews as a circuit breaker** for content on the path of being amplified. Services can apply filters to alert human moderators to content moving at high velocity, so this content be reviewed before going viral. This added frication helps to triage problematic content and can be useful as part of a service's crisis management process when policy shocks and crisis events like the Christchurch terrorist massacre rapidly surface. It's also important to acknowledge the limitations of human content reviews, including the potential for human bias and inconsistency.

- **Introducing additional friction** through design features, such as prompts to read an article linked before sharing it, or restricting how often content can be shared. Services like Twitter and Facebook have used nudges to prompt users to read an article before sharing it. These small interventions by services have the potential to make a significant difference.[12]

- **Labelling content as potentially harmful or hazardous**. Different communities experience harm differently, and content may be harmful to some people and not to others. Warnings, blurring and labels can be an effective way to help users avoid content they do not wish to see. Community groups should be consulted to make sure these warnings and labels are appropriate. Where content warnings are provided to some users and not others, consideration should be given to the data which informs these choices and the risk of bias.

- **Behavioural cues and prompts that can help users to establish positive patterns of behaviour**. Such features could help users reconsider posting harmful content, or to manage their time spent online. For example, some services prompt users posting captions that their algorithms may be deemed offensive before content is posted. Services can also incorporate messages to raise awareness of how their recommendations function and build users' algorithmic literacy.

- **Enhancing transparency** reporting and auditing practices. More information in the public domain, through researchers and experts or regulators, helps to improve on future interventions and regulations, and ultimately builds trust with users. Services can also benefit by engaging in international multi-stakeholder bodies developing technical standards for the ethical use and deployment of recommender systems.

- **Curating recommendations so they are age appropriate**. For example, Instagram by default filters out sensitive content that users under 18 can be recommended.

- **Parental controls** can give parents and carers visibility over what young people are viewing and to set restrictions over their online use. For example, Instagram's Family Centre allows parents and carers to manage time spent on the platform, see followers, and be notified when a child reports content.

Safety by Design

Advice for users

Many people find recommender systems a valuable way to discover new content, products and services which align with their interests. While there are numerous benefits, there are also risks, including the potential to be served with harmful content.

It can help to be aware of the systems, processes and business models underlying how content is suggested. When a service is recommending content to keep users engaged, they may rely on passive behaviours to profile a user and their interests. For example, the amount of time spent hovering over a piece of content before scrolling past may have as much as an influence on the recommender system as active behaviours such as liking a video.

Some services have introduced features that empower users to have some influence over how content is recommended. This could be through switching to a chronological or other type of feed as opposed to a service's main algorithmically generated feed.

Services may also offer the ability to flag specific content to be excluded from future recommendations. Users can experiment with these features. Features either installed by default or available through third-party apps can also help to monitor and restrict time spent online and avoid the potentially harmful mental health effects from excessive use, also known as 'doomscrolling' or 'doomsurfing'.

Further information and downloadable resources for parents and carers about screen time for children and advice about balancing time online for young people can be found on eSafety's website.

It can be helpful to understand that users play a crucial role in recommender systems. As such, users can act to shape and refine their recommendations, for example by intentionally not clicking on content they do not want to see.[13] People can also benefit from reflecting on the recommendations their algorithm produces and how that content makes them feel. This self-exploration can help people make active choices in shaping their recommender systems.

Recommender systems can contribute to and exacerbate the personal harms eSafety investigates, including the spread of illegal and restricted online content, cyberbullying of children, serious adult cyber abuse and image-based abuse. Where complaints are about harmful that has been publicly posted on a service, recommender systems risk amplifying this content and heightening the negative impacts experienced by victims.

Whether or not a recommender system is involved, if someone seriously abuses or harms you online (or someone in your care) there are several steps you can take:

1. Collect as much evidence as possible — for example, by taking screenshots (unless the shots show nude or sexual content or conduct of someone under 18 years old).
2. Report the abuse and the individual or account to the service. In Australia, you can also report seriously harmful content to eSafety (https://www.esafety.gov.au).
3. Prevent further contact once you have collected evidence – you can use in-app functions to ignore or mute them, or to block them after you have collected evidence.
4. Get more help – experiencing online abuse can be distressing, so you may find it helpful to talk with someone at an expert counselling and support service.

In Australia, scams can also be reported to Scamwatch, a website run by the Australian Competition and Consumer Commission (ACCC).

Digital Platform Regulators activities

The Digital Platform Regulators Forum (DP-REG) comprises representatives from the Australian Competition and Consumer Authority (ACCC), the Australian Communications and Media Authority (ACMA), the eSafety Commissioner (eSafety) and the Office of the Australian Information Commissioner (OAIC).

In June 2022, the heads of the four DP-REG members met and agreed on a collective set of priorities for 2022–23. This includes a focus on the impact of algorithms, enhancing transparency of digital platform activities and how they are protecting users from harms, and collaboration and capacity building between members.

DP-REG includes a sub-committee that is specifically examining algorithms and their impact.

eSafety will continue to support the DP-REG activities on algorithms and will update this position statement from time to time as DP-REG outputs and Australian Government considerations regarding algorithms progress.

Acknowledgements

eSafety acknowledges the contribution made by experts in sharing their insights on recommender systems with eSafety. In particular, we thank the following experts:

Louisa Bartolo, QUT Digital Media Research Centre and ARC Centre of Excellence for Automated Decision-Making and Society

Jenny Brennan, Ada Lovelace Institute

Dr Jonathan Bright, The Alan Turing Institute

Professor Jean Burgess, QUT Digital Media Research Centre and ARC Centre of Excellence for Automated Decision-Making and Society

Yi-Ling Chung, The Alan Turing Institute

Dr Ariadna Matamoros Fernandez, QUT Digital Media Research Centre and ARC Centre of Excellence for Automated Decision-Making and Society

Angus R Williams, The Alan Turing Institute

Dr Elizabeth Williams, ANU College of Engineering and Computer Science

We also extend our thanks to other academics who helped us to produce this paper.

[1] Borderline content, fringe or grey content means content that comes close to infringing a platform's terms of service or community guidelines but does not constitute a breach.

[2] Jonathan Bright, *Explaining the Emergence of Political Fragmentation on Social Media: The Role of Ideology and Extremism*, Journal of Computer-Mediated Communication, Volume 23, Issue 1, January 2018, Pages 17–33; Jonathan Bright, Nahema Marchal, Bharath Ganesh, Stevan Rudinac, *How Do Individuals in a Radical Echo Chamber React to Opposing Views? Evidence from a Content Analysis of Stormfront*, Human Communication Research, Volume 48, Issue 1, January 2022, Pages 116–145.

[3] Victoria Jaynes and Izzy Wick, *Risky by Design: Recommendation Systems*, 5Rights Foundation, 2022.

[4] Max Fisher, Amanda Taub, *On YouTube's Digital Playground, an Open Gate for Pedophiles*, The New York Times, June 3, 2019.

[5] North London Coroner's Service, *Regulation 28 Report to Prevent Future Deaths*, 13 October 2020. https://www.judiciary.uk/wp-content/uploads/2022/10/Molly-Russell-Prevention-of-future-deaths-report-2022-0315_Published.pdf

[6] Kari Paul, *Families sue TikTok after girls died while trying 'blackout challenge'*, The Guardian, 6 July 2022.

[7] Russell Brandom, *Inside Elsagate, the conspiracy-fueled war on creepy YouTube kids videos*, The Verge, 2017.

[8] eSafety Commissioner. *Protecting voices at risk online*. August 2020. esafety.gov.au/diverse-groups/protecting-voices-risk-online

[9] Angela Y. Lee, Hannah Mieczkowski, Nicole B. Ellison, and Jeffrey T. Hancock, *The Algorithmic Crystal: Conceptualizing the Self through Algorithmic Personalization on TikTok*, 2022 ACM Conference on Computer-Supported Cooperative Work and Social Computing, June 2022.

[10] Ada Lovelace Institute, *Technical methods for the regulatory inspection of algorithmic systems in social media platforms*, 2021.

[11] Francis Fukuyama, Barak Richman, Ashish Goel, Marietje Schaake, Roberta R. Katz, Douglas Melamed, *Report of the Working Group on Platform Scale*, Stanford Cyber Policy Centre.

[12] In research conducted in 2020, Twitter found that nudging users to read articles before retweeting through a 'Headlines don't tell the full story' prompt resulted in 40% more users reading the article before retweeting. See Twitter Comms, September 25 2020. https://twitter.com/TwitterComms/status/1309178716988354561

[13] Angela Y. Lee, Hannah Mieczkowski, Nicole B. Ellison, and Jeffrey T. Hancock, *The Algorithmic Crystal: Conceptualizing the Self through Algorithmic Personalization on TikTok*.

esafety.gov.au