Tech Trends Position Paper

Recommender Systems

May 2026

Australian Government

eSafetyCommissioner

eSafety.gov.au

Contents

Content warning: This paper discusses harms people may find distressing, including violent extremism and gender-based violence. Please take care when reading and consider whether this paper is right for you at this time.

Executive summary

Recommender systems and the algorithms that power them are integral to the architecture of the internet. These systems affect people's experiences in many ways that can be both helpful and harmful.

Social media platforms use recommender systems to show users posts, ads and accounts to follow, typically based on what a user is most likely to engage with. Search engines recommend content in response to search terms, as well as other user information, such as location. Streaming platforms recommend content, dating apps recommend matches and video games recommend in-game content based on many factors, including previous behaviour or popularity across a service, while AI chatbots use algorithms to generate responses to prompts or queries.

This can often be beneficial, allowing users to see relevant, high-quality and engaging content instead of spam. However, when recommender systems do not prioritise user safety, they can facilitate the spread and reach of harmful content, amplify the impact of sensitive content and enable other risky features, such as infinite scroll and autoplay, which can encourage excessive engagement.

Recommender systems are also opaque. Researchers, regulators and users often have limited visibility into how they work and how they shape users' experiences, which raises issues relating to transparency and accountability, as well as user autonomy and choice.

The interconnected issues associated with recommender systems require a layered and multidimensional approach. Singular solutions are often simplistic and can create unintended challenges. Instead, holistic approaches provide greater flexibility, the opportunity for meaningful and balanced safety improvements across the entire digital ecosystem, and the ability to reflect and respond to broader social contextual factors.

This position statement from the eSafety Commissioner (eSafety) explores the design of recommender systems, how they can amplify and compound harm online, and how systems should be designed to put user safety first and foremost. We consider how the overall design of recommender systems, rather than specific types of content, impacts users, and how the underlying social issues recommender systems draw upon and exacerbate can also shape harm. We then explore possible harm mitigation strategies and their associated challenges.

Lastly, we outline what safer systems could look like, especially when underpinned by Safety by Design. We also touch upon the systemic reform of the upcoming duty of care, which the Government has committed to introducing under the Online Safety Act.

Background

This position statement

As an anticipatory regulator, a core part of eSafety's role is to stay aware of technological developments and anticipate the emerging safety risks and possibilities they present.

Our **Tech Trends and Challenges program** helps us understand existing and emerging online threats and opportunities, which informs and shapes our regulatory guidance, resources and advice to industry.

In December 2022, we published our position statement on recommender systems and algorithms. Since then, eSafety has implemented and started to enforce new regulatory frameworks, including the **Social Media Minimum Age** and the **Online Safety Codes and Standards**. Platform and user behaviour has also changed – social media feeds contain more recommended content instead of friend or followed content,[1] and content is increasingly generated or otherwise facilitated by artificial intelligence (AI). These changes affect how we understand and respond to the harms associated with recommender systems and algorithms.

While there are many harms associated with recommender systems, this paper focuses on those related to eSafety's remit. It also examines recommender systems that serve up content on social media, with references to search engines and AI chatbots.

As an online safety regulator, eSafety takes a harms-minimisation and risk-based regulatory approach. We also adopt a strengths-based and intersectional approach to understanding and responding to a person's online experience. This guides our framing and discussion of emerging technologies.

This updated position statement outlines current and emerging risks and harms relating to recommender systems, identifies opportunities, explores regulatory challenges, provides guidance on Safety by Design measures, and explains what this means for the online safety of all users and systems.

This position statement reflects eSafety's position as of May 2026.

[1] J Chhabra et al., 'Social Media and Youth Mental Health: Scoping Review of Platform and Policy Recommendations', *Journal of Medical Internet Research,* 2025, 27, https://www.jmir.org/2025/1/e72061.

Who we are

eSafety is Australia's independent regulator and educator for online safety.

Under the *Online Safety Act 2021* (Cth) (the Act), we coordinate government efforts to improve online safety for all Australians. We conduct research, provide education, and enforce regulatory schemes to combat online harm. As part of our multidimensional regulatory remit, we work with government agencies, businesses and organisations – throughout Australia and around the world – to make the internet a safer place.

The Act gives eSafety a range of powers to help protect Australians from harm and promote safer online experiences. These include regulatory schemes operated by eSafety:

- **Online Safety Codes and Standards** (Unlawful Material Codes and Standards and the Age-Restricted Material Codes), which contain mandatory and enforceable compliance measures for key sections of the online industry. They cover material that is either the most seriously harmful online content, such as child sexual exploitation material and pro-terror material, or material that is inappropriate for children, such as online pornography.
- The **Basic Online Safety Expectations** (the Expectations), which set out the Australian Government's expectations for certain services to take reasonable steps to keep users safe. eSafety is also empowered to compel information from certain services about their compliance with the Expectations and to publish transparency reports.
- The **Social Media Minimum Age** obligation (SMMA), which requires age-restricted social media platforms to take reasonable steps to prevent Australians under the age of 16 from having an account on their services.

The Act also includes four reporting schemes for the following types of material:

- **Adult cyber abuse**, which relates to seriously harmful material targeting Australian adults (18 years or older).
- **Child cyberbullying**, which relates to seriously harmful material targeting Australian children (under 18 years).
- **Image-based abuse**, which relates to the non-consensual sharing, or threat to share, intimate images or videos.
- **Illegal and restricted content**, which relates to **class 1** or **class 2** material.

How the Act applies to recommender systems and algorithms

The Act applies to recommender systems and algorithms in different ways depending on the regulatory schemes.

The **Basic Online Safety Expectations** (the Expectations) outline expectations relating to recommender systems.[2] Across several expectations, platforms are expected to take reasonable steps to ensure their platforms can be used safely, and to limit access to certain types of material. **Additional Expectation 8B** relates specifically to the use of recommender systems.

- The **Determination** and eSafety's **Regulatory Guidance** outline examples of reasonable steps platforms can take, including providing opt-in or opt-out measures for users, and designing algorithms to consider metrics such as authoritativeness or diversity.
- eSafety can require platforms to report on the steps they are taking to comply with the Expectations, helping to improve transparency.

The **Online Safety Codes and Standards** include obligations for sectors of the online industry to provide appropriate safety tools that allow users to limit access to certain material. Appropriate safety tools may include designing recommender systems to ensure they do not promote such material.[3] Additionally, Internet Search Engine Services have obligations that apply to recommender systems that provide or personalise search results.[4]

We combine our protective regulatory schemes with prevention initiatives and proactive and systemic change measures. As part of this holistic approach, we undertake a range of activities to improve digital literacy and better understand the evolving opportunities, risks and impacts of recommender systems.

For example, we are developing education and training programs to help people understand the risks of recommender systems and the tools they can use to reduce exposure to harmful content and manage their feeds.

[2] Including Expectations 6, 11 and 12.
[3] See: clause 7.2 of the Social Media Services (Core Features) Online Safety Code (Class 1C and Class 2 Material) and clauses 7.5 and 10.21 of Designated Internet Services Online Safety Code Class 1C and Class 2 Material).
[4] See: Internet Search Engine Services Online Safety Code (Class 1A and Class 1B Material) and Internet Search Engine Services Online Safety Code (Class 1C and Class 2 Material).

What are algorithms and recommender systems?

Recommender systems are central to how we use the internet and are integral to many online services.

They select, filter and personalise content and other items across online platforms, services and applications. They help users sort through vast amounts of data, discover new artists, friends, products, activities and ideas, and help businesses and creators reach new audiences.

This section provides a technical overview of algorithms and recommender systems. We discuss risks and associated harms on page 10.

Key terms

An **algorithm** refers to a set of computing instructions or rules followed to complete a specific task or solve a problem. Its instructions are based on many factors. In this context, algorithms assess pieces of content and rank them based on a formula that considers different signals and weightings.

Platforms generally rely on multiple algorithms, each designed for different content or functions, to create a unique, personalised platform experience for each user. The combination of these algorithms – plus business considerations, system-level constraints, design features, user controls and moderation policies – make up a **recommender system** that suggests relevant items or pieces of information to a user.

While an algorithm will follow a given set of computing instructions to provide a score for a piece of content, a recommender system takes the results from multiple algorithms, along with other data sources, to determine what content is recommended to a user.

In this paper, we use the term ‘recommender system’ to discuss the **overall system used by a platform**.

How do recommender systems work?

Each platform has its own purposes when recommending content and, as a result, will design the algorithms that power its recommender systems differently.[5]

Complex recommender systems, like the ones powering social media platforms, generally rely on a **neural network** – a type of machine learning model that mimics the human brain – to generate ranking scores which help decide which content to display to an end user. They do this by applying machine learning techniques to data held by online services to make predictions based on user attributes and patterns and then make recommendations based on these predictions. Social media algorithms often use these predictions to estimate how likely a user is to engage with content presented to them.

Users are also increasingly being recommended content online through **generative AI tools, such as AI chatbots and AI search summaries**. These tools use **natural language processing algorithms** to respond to user requests in a conversational manner. These algorithms are used to train computers to understand and process human language using techniques – such as sentiment analysis, keyword extraction and text summarisation – to analyse a user query and provide a conversational response.[6] This allows them to summarise information from multiple sources and answer multi-step queries.

Signals used by recommender systems

Modern recommender systems use many signals to inform recommendations. These signals can be sourced in several ways. Data may be provided **knowingly** by users or collected by **inferring** information about a user based on their behaviour.

Recommender systems use these data points to give each piece of content a specific **score,** often tailored to each user. This determines the order in which content is shown. Scores are also adjusted for other factors, such as avoiding repetition or limiting harmful content.

Data used to sort content can be grouped into two categories: **signals** and **predictions**.[7]

---

[5] Meta, *The AI behind unconnected content recommendations on Facebook and Instagram,* 2023, https://ai.meta.com/blog/ai-unconnected-content-recommendations-facebook-instagram/; TikTok, *How TikTok Recommends Content,* n.d, https://support.tiktok.com/en/using-tiktok/exploring-videos/how-tiktok-recommends-content**;** YouTube, *Recommendations on YouTube,* 2026, https://www.youtube.com/howyoutubeworks/recommendations/.

[6] IBM, *What is NLP (natural language processing)?,* n.d, https://www.ibm.com/think/topics/natural-language-processing.

[7] A Moehring et al., 'Better Feeds: Algorithms that put people first', 2025, *KGI Expert Working Group on Recommender Systems,* https://kgi.georgetown.edu/research-and-commentary/better-feeds/.

Signals range from observations about behaviour to qualitative data and explicit user controls,[8] such as:

- **Engagement or behavioural signals.** These include actions a user takes on an item, such as clicks, shares, likes, dwell time or purchases.
- **User responses to survey questions.** These can be questions about a specific piece of content, such as 'would you like to see more of this?', or more general questions about the user's experience on the platform.
- **Information about a piece of content.** This includes the type of media (such as text, video or photo) and often includes analysis of sentiment or topic.
- **User controls.** This includes settings a user engages to control what type of content appears to them. These are different for each platform but can include features such as blocking or muting, different feed options, such as 'for you' or 'discover' feeds and parental controls.
- **Profile data.** This can include a user's age, gender, religion, political affiliation, socioeconomic status, location, race and ethnicity.
- **Off-platform data.** Platforms can use cookies[9] to track a user's browsing activity outside the platform's suite of services.
- **Other contextual data.** This includes the date, time of day and location.

Platforms build on these signals by predicting the likelihood of certain outcomes or responses to content. Some major categories of **predictions** include:

- **Engagement predictions.** The likelihood that a user will engage or interact with a piece of content, based on existing data the platform has about the user, such as previous engagement signals and demographic data. This is commonly used in social media feeds.[10]
- **Content scores.** This is the score assigned to a piece of content based on the platform's priorities. For example, higher scores could be assigned to content posted by accounts known to be trusted news accounts, while lower scores could be given to content that comes close to, but does not breach, moderation policies.

[8] Moehring et al., 'Better Feeds: Algorithms that put people first'.
[9] Cookies are small files of information that a web server sends to the user's web browser to store for future use. These files allow browsers to remember information such as usernames and passwords, other URLs visited, security information and browsing habits.
[10] Moehring et al., 'Better Feeds: Algorithms that put people first'.

Much of this data relies on a combination of machine learning and fixed weights.[11] Traditional approaches to recommender systems assign fixed weights to specific predictions – for example, giving higher ranking scores to more recent content or to content where a user watches the full video – to determine a score for each piece of content.

Using **machine learning**, **neural networks** and **automated decision making** allows these systems to generate more complex, adaptable and personalised content rankings that can handle larger volumes of content.

Online platforms make decisions about optimising their recommender systems to achieve different purposes, which can have significant impacts on user experience. For example, a platform may seek to prioritise engagement, maximise time spent on its service, deliver recommendations that best meet users' needs, increase the diversity of content that users are exposed to, or pursue some combination of all of these. We discuss the consequences of these decisions in the following section.

[11] Fixed weights are parameters that remain the same during the training and deployment of an algorithm or recommender system. They are pre-determined and structural.

Exploring the risks, harms and impacts of recommender systems

This section outlines the unique presentations of risks and harms associated with recommender systems, including how they are affected by context and specific design choices, and how these extend beyond the impacts of individual pieces of content.

Why recommender systems can amplify harms online

Recommender systems are built to prioritise certain content. When the decisions and policies shaping their design do not put users' best interests first, they can extend the reach of harmful content and amplify its negative impact.

Recommender systems also draw on and exacerbate underlying social issues, such as inequality, discrimination, prejudice and oppression. This includes harms like sexism, racism, ableism, homophobia, ageism and transphobia, often in intersecting and cumulative ways. It is therefore important to understand how recommender systems interact with harms that already exist in society, including:

- How content users share and engage with online is often rooted in broader societal issues or polarising and emotionally charged debates – such as political issues – rather than simply a service's recommender system.[12] A user's personal experience also impacts the potential harm of this content.
- How recommender systems amplify, extend, or help spread this content, which is shaped by underlying platform design and features, and what weights and metrics are used.
- How harmful content can be recommended at scale, creating and extending harm in a way that is greater than the sum of its parts.

Since recommender systems can also create and exacerbate harms at a societal level,[13] it is important to take a holistic view of them. This includes their benefits and risks, broader uses and the complex interconnected ecosystems they operate within.

---

[12] J Hakkarainen and L Savolainen, 'Individual choice, collective effects: recommender systems, law by design, and the DSA's double choice architecture', *Information, Communication & Society,* 2025, https://doi.org/10.1080/1369118X.2025.2595663.

[13] Hakkarainen and Savolainen, 'Individual choice, collective effects: recommender systems, law by design, and the DSA's double choice architecture'.

These dynamics require a targeted response that differs from tackling individual pieces of content.

Context matters

The impact of content, and the severity and type of harm a user may experience because of content recommendations, varies based on the user's context and circumstances. These may include intersectional factors specific to the individual, as well as the features and functions of the platform itself. For example, content that promotes self-harm is likely to present a greater risk to someone already experiencing, or who has, or is likely to experience, mental ill-health. Similarly, content shown with a visible 'like' count may affect users in different ways.[14]

Recommender systems rarely consider nuance or context.

Recommender systems often lack nuance or context and do not always distinguish whether content about a sensitive topic is harmful or intended to reduce harm. This can mean that users seeking harm-reduction content can be exposed to material that advocates, promotes, or reshares the very harm they are trying to reduce their exposure to.

For example, some platforms recommend content that promotes disordered eating or documents a person's relapse to users who engage with eating disorder recovery content.[15] While some users with lived and living experience of eating disorders may find relapse content helps with their recovery, the same content may negatively affect other users with lived and living experience.[16] Recommender systems are also not always responsive when a user wants to move away from engaging with content that is harmful to them.

[14] A Voggenreiter et al., 'The Role of Likes: How Online Feedback Impacts Users' Mental Health', *WEBSCI '24: Proceedings of the 16th ACM Web Science Conference*, 2024, https://doi.org/10.1145/3614419.3643995.

[15] Reset Australia, *Not just algorithms: Assuring user safety online with systemic regulatory frameworks,* 2024, https://apo.org.au/node/326122.

[16] J Golbeck, 'Recommender System-Induced Eating Disorder Relapse: Harmful Content and the Challenges of Responsible Recommendation', ACM Trans. Intelligent Systems Technology 2025, 16(1), https://dl.acm.org/doi/pdf/10.1145/3675404; C F Scott et al., 'Trauma-informed social media: Towards solutions for reducing and healing online harm', *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems,* 2023, https://dl.acm.org/doi/epdf/10.1145/3544548.3581512.

Impacts and harms are cumulative

Recommender systems can compound harm by recommending large volumes of harmful content.[17]

Cumulative harm occurs when a person is harmed by repeated negative effects. These harms may or may not, on their own, be considered harmful, but become increasingly harmful when they are multiple and successive. When recommender systems push content onto a user's feed at scale, the cumulative effect can be far more damaging than any individual post, creating deeper and longer lasting harm.[18]

Consequences of design choices

The design of recommender systems, including how and why they prioritise content, has a range of consequences that can facilitate and amplify risk and harm.

Facilitating the spread and reach of harmful content

One of the most common concerns with recommender systems is the role they play in facilitating the spread and reach of harmful content. A key driver of this risk is the way social media services optimise their recommender systems for greater engagement.

Social media platforms overwhelmingly **prioritise engagement** when sorting and presenting content algorithmically. They rely on signals such as dwell time,[19] likes, shares and comments, as well as inferred predictions, to maximise the time a user spends on their service. This assumes that engagement means the user wants to see more of that content.[20]

---

[17] L Bartolo and A Matamoros Fernandez, 'Online Harm', *ISP-platform governance terminologies essay series,* 2023, 27, https://eprints.qut.edu.au/241428/.

[18] Bartolo and Matamoros Fernandez, 'Online Harm'; P Madriaza et al., 'Exposure to hate in online and traditional media: A systematic review and meta-analysis of the impact of this exposure on individuals and communities', *Campbell systematic reviews,* 2024, 21(1), https://pubmed.ncbi.nlm.nih.gov/39822240/.

[19] Dwell time refers to how long a user spends interacting with or looking at a specific piece of content.

[20] Moehring et al., 'Better Feeds: Algorithms that put people first'; S Milli et al., 'Engagement, user satisfaction, and the amplification of divisive content on social media', *PNAS nexus,* 2025, 4(3), https://academic.oup.com/pnasnexus/article/4/3/pgaf062/8052060?login=false.

However, these metrics are often poorly aligned with pro-social values[21] or with users' intentions and long-term desires.[22] Engagement-based recommender systems **rarely distinguish between positive and negative engagement** and users often interact with content they do not find valuable.[23]

When a user encounters content and reacts negatively, a recommender system may still treat that response as a signal to recommend more of the same content. Content that triggers a negative reaction may also be pushed to other users. This means harmful content that provokes a strong reaction can be amplified across a platform.

At an individual level, encountering harmful content online, including violent or extreme material, can be highly distressing. At a broader societal level, amplifying content that promotes inequality, discrimination and oppression can normalise prejudice or hate and contribute to a broader undermining or destabilising of social cohesion. This may also contribute to radicalisation towards terrorism and violent extremism.[24]

People who actively seek out extreme views online can find them regardless of how recommender systems rank content. However, recommender systems also play a role in distributing violent, extreme or harmful content to users who are not intentionally searching for it.[25]

The 'manosphere'

The 'manosphere' is a network of loosely interconnected content creators who claim to address men's struggles in areas such as dating, fitness and gaming, but often replicate misogynistic worldviews[26] and harmful versions of masculinity.[27]

---

[21] Moehring et al., 'Better Feeds: Algorithms that put people first'.
[22] Milli et al., 'Engagement, user satisfaction, and the amplification of divisive content on social media'.
[23] Milli et al., 'Engagement, user satisfaction, and the amplification of divisive content on social media'.
[24] ISD, 'Misogynistic Pathways to Radicalisation: Recommended Measures for Platforms to Assess and Mitigate Online Gender-Based Violence', 2023, https://www.isdglobal.org/publication/misogynistic-pathways-to-radicalisation-recommended-measures-for-platforms-to-assess-and-mitigate-online-gender-based-violence/.
[25] Australian Institute of Criminology, *Exposure to and sharing of fringe or radical content online*, 2024, https://www.aic.gov.au/publications/tandi/tandi705.
[26] B Milne and C R Baker, 'From 'villains' to 'idols': exploring teenage boys' conflicting attachments to manospheric masculinities', *Gender and Education,* 2025, https://www.tandfonline.com/doi/pdf/10.1080/09540253.2025.2568407; K Regehr, C Shaughnessy, M Zhao, and N Shaughnessy, 'Safer Scrolling: How algorithms popularise and gamify online hate for young people', *UCL IOE, University of Kent,* 2024, https://www.ascl.org.uk/ASCL/media/ASCL/Help%20and%20advice/Inclusion/Safer-scrolling.pdf; UN Women, *What is the manosphere and why should we care?*, 2025, https://www.unwomen.org/en/articles/explainer/what-is-the-manosphere-and-why-should-we-care.
[27] F O'Rouke, C Baker, and D McCashin, 'Addressing the impact of Masculinity Influencers on Teenage Boys', 2025, https://www.dcu.ie/antibullyingcentre/addressing-impact-masculinity-influencers-teenage-boys.

Manosphere content often blames women, gender-diverse and LGBTIQ+ people, as well as migrants,[28] for the problems boys and young men face. This perpetuates the harmful narrative that feminism and gender equality have come at the expense of men's rights.

While the early popular manosphere creators were often straight white men, more recently men of colour and 'tradwife'[29] influencers have also become popular.[30]

Recommender systems play a foundational role in spreading manosphere content.[31] **Boys and young men are consistently recommended manosphere content, regardless of whether they seek it out**.[32] While self-help and advice content can offer real benefits, it often serves as a gateway to, and a cover for, increasingly extreme and malicious manosphere content. This material often deliberately plays on boys' and young men's feelings of insecurity, resentment and shame, as well as their desires for identity, belonging and validation. Repeated exposure can create an echo chamber effect that distorts how common and prominent these ideas appear to be.

eSafety's research into young men online found that boys and young men often engage with this content more critically than commonly presumed.[33] Many expressed critical awareness about how algorithms steer them towards unwelcome or harmful content.[34] The fact that many boys and young men feel uncomfortable about manosphere content highlights the role recommender systems play in its spread and promotion.

Recommender systems are vulnerable to manipulation

Recommender systems that place outsized emphasis on engagement may encourage users to make their content more inflammatory, extreme or polarising to **boost engagement** on their posts for reasons including financial gain or wider influence.[35]

---

[28] O'Rouke et al, 'Addressing the impact of Masculinity Influencers on Teenage Boys'.
[29] A 'tradwife' (short for 'traditional wife') is an internet subculture of women who promote traditional family structures and values, including prioritising homemaking and raising children. Tradwife content often also advocates for problematic gender roles, such as women's submissiveness to their husbands, rejection of women's financial autonomy and – in some cases – the promotion of white nationalist values. See I Richards, C Jones, and V Trott, 'Neoliberal capitalism and the political aesthetics of tradwife imagery', *Distinktion: Journal of Social Theory,* 2025, 27(1), https://www.tandfonline.com/doi/full/10.1080/1600910X.2025.2550259.
[30] V Gerrand, D Ging, J M Roose, and M Flood, 'Mapping the Neo-Manosphere(s): New Directions for Research', *Men and Masculinities*, 2025, *28*(5**),** https://doi.org/10.1177/1097184X251350277; D Procope Bell, '"Pick-Me" Black women: tactical patriarchal femininity in the Black manosphere', *Feminist Media Studies*, 2024, 24(8), https://doi.org/10.1080/14680777.2023.2262163**.**
[31] Gerrand et al, 'Mapping the Neo-Manosphere(s): New Directions for Research'; Regehr et al, 'Safer Scrolling: How algorithms popularise and gamify online hate for young people'.
[32] C Baker, D Ging, and M Brandt Andreasen, 'Recommending toxicity: the role of algorithmic recommender functions on YouTube Shorts and TikTok in promoting male supremacist influencers', 2024, https://www.dcu.ie/antibullyingcentre/recommending-toxicity-summary-report.
[33] eSafety, *Young men online,* 2023, https://www.esafety.gov.au/research/young-men-online.
[34] eSafety, *Young men online*.
[35] Milli et al., 'Engagement, user satisfaction, and the amplification of divisive content on social media'.

These systems can also be manipulated through tactics such as rage-bait content,[36] coordinated networks or fake accounts[37] and targeted hashtags.[38] Malicious users can buy fake followers and use bots, including bot networks,[39] to inflate likes, comments, shares and replies at low cost.[40]

This can artificially boost visibility and allow harmful or misleading material to spread further and faster. Platform design choices can unintentionally amplify these behaviours by rewarding engagement over safety.

Impact of AI-generated material

AI generation tools now enable users to rapidly produce large volumes of content, which malicious actors can exploit to flood platforms, evade moderation and create the illusion of widespread engagement.

These tools have already been used to amplify racist and extremist narratives, including through AI-generated images and videos that fuel misinformation and social division, reinforce echo chambers and distort information ecosystems.[41]

Social media platforms generally do not distinguish between AI-generated and user-generated content when making recommendations. Together, these technologies can amplify each other in ways that substantively reshape users' information ecosystems.[42] This exacerbates cumulative and contextual harms.

---

[36] K Scott, ABC, *The 'vicious cycle' of rage bait and how to avoid it*, 2025, https://www.abc.net.au/news/2025-05-01/understanding-rage-bait-and-how-to-spot-it/105179784.

[37] F Menczer, The Conversation, *Swarms of AI bots can sway people's beliefs – threatening democracy,* 2026, https://theconversation.com/swarms-of-ai-bots-can-sway-peoples-beliefs-threatening-democracy-274778.

[38] ISD, 'Misogynistic Pathways to Radicalisation: Recommended Measures for Platforms to Assess and Mitigate Online Gender-Based Violence'.

[39] Social media bots are automated programs that simulate human engagement on social media platforms. Bots often act as part of coordinated networks, flooding social media with similar content and reposting content from one bot to another. Bot networks can amplify disinformation and create a false sense of public opinion on an issue. For example, researchers found that more than half of the Twitter accounts discussing COVID-19 were bots. AI models make it much easier and cheaper for malicious agents to create and manage these accounts at scale.

[40] D Nevado Catalan et al., 'An analysis of fake social media engagement services', *Computers & Security* 2023, 124, https://www.sciencedirect.com/science/article/pii/S0167404822004059.

[41] N Dirga, AAP, *Foreign Facebook page using AI-generated women to stir immigration outrage*, AAP, https://www.aap.com.au/factcheck/foreign-facebook-page-using-ai-generated-women-to-stir-immigration-outrage/?_bhlid=e4183ee326202571828aed3995d989878766890d; J R Kunst et al., 'Intelligent Systems, Vulnerable Minds: A Framework for Radicalization to Violence in the Age of AI', *Personality and Social Psychology Review,* 2026, https://journals.sagepub.com/doi/pdf/10.1177/10888683261430089; J Wilkins, Futurism, *Racists Are Using AI to Spread Diabolical Anti-Immigrant Slop*, 2025**,** https://futurism.com/future-society/racists-anti-immigrant-slop; M Workman, M Martino, and L Carter, ABC, *Foreign Facebook accounts using AI Pauline Hanson to manipulate Australians*, 2026, https://www.abc.net.au/news/2026-03-11/foreign-fake-news-pauline-hanson-one-nation/106436702.

[42] J Tidy, BBC, *AI 'slop' is transforming social media - and a backlash is brewing,* 2026, https://www.bbc.com/news/articles/c9wx2dz2v44o.li**.**

Broader impacts on wellbeing

Recommender systems pose a range of risks to user wellbeing, including:

- **Distorted reality**. Content rabbit holes and echo chambers can reinforce harmful attitudes and unrealistic expectations about life, relationships and success.[43] Recommender systems may also contribute to repeated exposure to content that promotes 'ideals' of body types and beauty stereotypes.[44]
- **Dependency.** Features such as infinite scrolling and autoplay, which are enabled by recommender systems, as well as push notifications, feedback loops and rewards systems, may encourage compulsive use and contribute to excessive time online.[45]
- **Risky interactions.** Young users may be recommended individuals or communities that are unsafe, including through friend or follower suggestions that pressure children to interact with potentially dangerous adults.[46]
- **Access to age-inappropriate content.** Recommender systems may suggest content that is appropriate for adults but harmful to children who are not developmentally ready for it, such as violent or sexually explicit material.[47]

[43] Milne and Baker, 'From 'villains' to 'idols': exploring teenage boys' conflicting attachments to manospheric masculinities'; R O'Connell and N A Daruwala, 'Milestones and mindsets: how social media shapes young adults' expectations and emotional well-being', *Journal of Social Media Research*, 2025, 2(5), https://jsomer.org/index.php/pub/article/view/56/47.

[44] R C Bonfanti, 'The association between social comparison in social media, body image concerns and eating disorder symptoms: A systematic review and meta-analysis', *Body image,* 2025, 52 https://www.sciencedirect.com/science/article/pii/S1740144524001633.

[45] American Psychological Association (APA), 'Potential Risks of Content, Features, and Functions: A closer look at the science behind how social media affects youth', 2024, https://www.apa.org/topics/social-media-internet/youth-social-media-2024; M Wolgast, H Adler, and S N Wolgast, 'Motives for social media use in adults: associations with platform-specific use, psychological distress, and problematic engagement', *Journal of Social Media Research*, 2025, 2(3), https://jsomer.org/index.php/pub/article/view/45/31.

[46] M Scherer and K Tiffany, The Atlantic, *How Meta Executives Talked About Child Safety Behind the Scenes*, 2026, https://www.theatlantic.com/technology/2026/02/meta-child-safety-documents-instagram/686163/.

[47] APA, 'Potential Risks of Content, Features, and Functions: A closer look at the science behind how social media affects youth'; J Robinson et al, 'How do Australian social media users experience self-harm and suicide-related content? A National cross-sectional survey comparing young people and adults', *BMC Public Health,* 2026 https://doi.org/10.1186/s12889-025-25646-0.

Young people's wellbeing

Features that are designed to keep users engaged for longer periods of time, such as infinite scroll, may encourage passive social media use, which some evidence suggests is associated with higher mental health symptoms.[48]

When not part of a broader balanced approach, social media use can get in the way of everyday relationships and experiences, and other important behaviours.[49] A 2023 survey by the Black Dog Institute found that 50% of young people said screen use was displacing other activities in their lives. For those that used a screen-based device before going to sleep, 45% reported going to sleep later than they wanted to, due to the device.[50]

In a 2024 report from Project Rockit, young people reported both positive and negative experiences from social media algorithms. However, some young people reported 'feeling trapped' by endless consumption of content, not enjoying content but watching it anyway, and feeling bad or regretful for 'bed rotting' and 'doomscrolling' late into the night.[51] The report also highlights the features missing from current platforms, which young people want, to allow them greater agency over their own experiences. This includes turning on or off features like infinite scroll, actively shaping recommendation preferences and tools to limit their own usage.

In March 2026, the Government updated the **legislative rules** to further define 'age-restricted social media platforms' for the purposes of the Social Media Minimum Age obligation, to include where services adopt **recommender features** and associated **logged in features** including endless feeds, feedback features and time-limited features. The Government stated this was to 'ensures the law remains focused on features that drive addictive behaviour and pose the greatest risk of harm to young people'.

[48] Black Dog Institute, 'Adolescent screen use and mental health: Summary of findings from the Future Proofing Study', 2024, https://www.blackdoginstitute.org.au/research-projects/teens-screens-adolescent-mental-health-screen-use/.
[49] Black Dog Institute, 'Adolescent screen use and mental health: Summary of findings from the Future Proofing Study'.
[50] Black Dog Institute, 'Adolescent screen use and mental health: Summary of findings from the Future Proofing Study'.
[51] Project Rockit, 'Shaping Our Feeds, Young people's experiences of social media algorithms', 2024, https://www.projectrockit.com.au/shaping-our-feeds.

Current harm mitigation strategies and challenges

Users and recommender systems interact in unique, interconnected and complex ways. Content is recommended to a user based on both the choices platforms make and how users interact with the service. The way that content is received, and the impact it has, also depends on the user and their activity. While there is no simple way to mitigate these harms, there are several steps industry can, and should, take to reduce them.

This section outlines the different strategies currently used to address harms associated with recommender systems. It also explores the limits of these approaches and what can be learnt in refining these strategies going forward.

Moderating the content

Recommender systems and content moderation are closely interrelated.

Content moderation often refers to the process of reviewing whether a piece of content complies with a platform's policies or community standards.[52] Content that breaches these policies may be deleted or removed. However, content that comes close to breaching them, but does not meet the threshold for removal, can be downranked or deprioritised by the platform's recommender system, limiting its reach. This is often described as 'borderline content'.

Recommender systems routinely exercise content moderation decisions to limit the reach of 'borderline content'. These decisions are often not explained to users, and are difficult to observe, measure or report. This creates transparency and accountability issues.

[52] Trust and Safety Professional Association, *What is content moderation?,* n.d, https://www.tspa.org/curriculum/ts-fundamentals/content-moderation-and-operations/what-is-content-moderation/.

Moderating content on large social media platforms is complex and has its own risks and shortcomings. Moderation alone cannot address the harms associated with recommender systems:

- Content moderation increasingly relies on automated decision-making with limited human oversight.[53] Even state-of-the-art detection systems cannot reliably infer user intent,[54] or quickly adapt to new trends[55] and coded language.[56] This means automated detection software is often not enough, even with some human oversight, to enforce platform policies.[57]
- User reporting of harmful material is often low. Contributing factors include retraumatising and confusing processes, inconsistent and uninformed policies, and poor information sharing across platforms to enable coordinated action.[58]

Transparency

The way online services use recommender systems is often opaque and can change frequently in response to a range of factors, including user input, commercial pressures, regulation and public sentiment. Without greater transparency, users cannot fully understand their online experiences, and regulators and researchers cannot properly scrutinise platforms' design and decision-making to hold them accountable when recommender systems cause harm.

---

[53] ABC, *TikTok slashes hundreds of jobs to help boost AI-assisted content moderation*, 2024, https://www.abc.net.au/news/2024-10-12/tiktok-slashing-jobs-to-boost-ai-content-moderation/104465606; K Bell, Engaget, *Meta will move away from human content moderators in favor of more AI*, 2026, https://www.engadget.com/social-media/meta-will-move-away-from-human-content-moderators-in-favor-of-more-ai-183000435.html; T Bennet, AFR, *How TikTok, YouTube and X police their platforms*, 2025, https://www.afr.com/technology/how-tiktok-youtube-and-x-police-their-platforms-20250108-p5l2ra; D Kayyali, Tech Policy Press, *Meta's Content Moderation Changes are Going to Have a Real World Impact. It's Not Going to be Good*, 2025, https://www.techpolicy.press/metas-content-moderation-changes-are-going-to-have-a-real-world-impact-its-not-going-to-be-good/.

[54] X Wang et al., 'The unappreciated role of intent in algorithmic moderation of abusive content on social media', *Harvard Kennedy School Misinformation Review 1*, 2025, 6(3), https://misinforeview.hks.harvard.edu/article/the-unappreciated-role-of-intent-in-algorithmic-moderation-of-abusive-content-on-social-media/.

[55] M Mehta and F Guinchiglia, 'Understanding Gen Alpha's digital language: Evaluation of LLM safety systems for content moderation', *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency,* 2025, https://dl.acm.org/doi/full/10.1145/3706598.3713998

[56] Mehta and Guinchiglia, 'Understanding Gen Alpha's digital language: Evaluation of LLM safety systems for content moderation'.

[57] F dos Santos et al., EMILDAI, *Decisions Without Reasons: The Achilles Heel of Social Media Moderation,* 2025, https://emildai.eu/decisions-without-reasons-the-achilles-heel-of-social-media-moderation; A Wiesner, S Schafer, S Lecheler, 'Navigating grey areas of content moderation Professional moderators' perspectives on uncivil user comments and the role of (AI-based) technological tools', *New Media and Society,* 2025, 27(3), https://journals.sagepub.com/doi/full/10.1177/14614448231190901.

[58] S Hakimi, Tech Policy Press, *Tools for reporting online violence are broken. Here's how to fix them,* 2025, https://www.techpolicy.press/tools-for-reporting-online-violence-are-broken-heres-how-to-fix-them; V Vilk and K Lo, PEN AMERICA, *Shouting into the void: why reporting abuse to social media platforms is so hard and how to fix it,* 2023, https://pen.org/report/shouting-into-the-void.

Transparency has become more complicated as recommender systems have become more individualised. They are less focused on recommending content from friends and followed users and more focused on algorithmically selected content.

Discoverability is also becoming increasingly reliant on algorithmic sorting instead of other metrics, such as follower count.[59]

This makes transparency more difficult because it is harder to identify patterns across users, especially for those outside the platform ecosystem, such as regulators.

There are also challenges for the external oversight of transparency initiatives. For example, researchers have found that boys and young men are increasingly exposed to harmful manosphere content that is reuploaded by anonymous accounts, rather than specific influencers, and surfaced through recommendations. This highlights the role algorithms play in amplifying exposure to harmful content.[60]

Similarly, while transparency reporting under the European Union's Digital Services Act has made important information available and improved transparency, definitions and collection methods are not consistent across platforms. This has led to inconsistencies in risk assessment and audit processes.[61]

Greater transparency is critical to holding online services accountable for the impact of their design choices. For transparency to be truly useful, there needs to be a common understanding of key terms and expectations, as well as minimum standards for the data, evidence and documentation to be included in audit and risk assessment processes. This would improve public understanding, strengthen regulatory oversight and support researchers to monitor and report on platform safety practices.

---

[59] A Silberling, TechCrunch, *Social media follower counts have never mattered less, creator economy execs say,* 2025, https://techcrunch.com/2025/12/29/social-media-follower-counts-have-never-mattered-less-creator-economy-execs-say/; Chhabra et al, 'Social Media and Youth Mental Health: Scoping Review of Platform and Policy Recommendations'.

[60] Milne and Baker, 'From 'villains' to 'idols': exploring teenage boys' conflicting attachments to manospheric masculinities'.

[61] P Chapman, Tech Policy Press, *From ambiguity to accountability: Analyzing recommender system audits under the DSA,* 2025, https://www.techpolicy.press/from-ambiguity-to-accountability-analyzing-recommender-system-audits-under-the-dsa/; S Groesch, et al., 'Big data, small answers: How the DSA transparency database falls short of its regulatory objectives', *Telecommunications Policy,* 2026, https://www.sciencedirect.com/science/article/pii/S0308596125001855#sec2.

User controls

Many online platforms give users content controls to tailor the content recommended to them. These commonly include options to hide posts, block, mute and reset recommendations.[62] More recently, some platforms have also introduced topic-level feedback to help tailor recommender systems.[63] These tools can be particularly useful because they help reduce exposure to the kinds of content that present a greater risk of harm based on an individual's circumstances. These tools are much needed, with 7 in 10 adults frequently seeing content they consider harmful.[64]

However, content control tools have seen limited uptake among users.[65] This could be due to a range of factors, including:

- **Interface design.** Changing where users access certain features can have knock-on effects for other tools. For example, research has found that making a 'hide post' tool easier to find increased the number of posts being hidden, but also decreased use of other content control features such as 'block' and 'report'.[66]
- **Unclear effects on user experience.** Many users are reluctant to use content tools. Some worry they will miss popular content seen by their friends[67] and others are sceptical about whether content control is effective.[68]

Improving the accessibility and discoverability of controls can increase their uptake and positive reception among users.[69] Content tools should be easy to find, and users should be given clear information about how they will influence content recommendations.

---

[62] Meta, *Reshape your Instagram with a recommendation reset,* 2024, https://about.fb.com/news/2024/11/introducing-recommendations-reset-instagram/; YouTube, *Recommendations,* n.d, https://www.youtube.com/howyoutubeworks/recommendations.
[63] Meta, *Control your threads feed with new dear algo feature,* 2026, https://about.fb.com/news/2026/02/threads-dear-algo/; TikTok, *TikTok Help: Manage Topics*, n.d, https://support.tiktok.com/en/account-and-privacy/account-privacy-settings/manage-topics.
[64] Ofcom, 'User empowerment and content control tools: Testing how design affects people's choices', 2025, https://www.ofcom.org.uk/online-safety/safety-technology/user-empowerment-and-content-control-tools-testing-how-design-affects-peoples-choices.
[65] In a study by Ofcom, only 26% of surveyed users said that they had used content controls, see: Ofcom, 'Behavioural insights to empower social media users: Testing tools to help users control what they see', 2024, https://www.ofcom.org.uk/online-safety/safety-technology/behavioural-insights-to-empower-social-media-users.
[66] Ofcom, 'User empowerment and content control tools: Testing how design affects people's choices'.
[67] Ofcom, 'User empowerment and content control tools: Testing how design affects people's choices'.
[68] Ofcom, 'User empowerment and content control tools: Testing how design affects people's choices'.
[69] Moehring et al., 'Better Feeds: Algorithms that put people first'.

Alternatives to engagement-based sorting

Different inputs and goals for recommender systems can lead to a range of outcomes, both positive to negative.

Reverse chronological or non-personalised feeds have become a commonly proposed alternative to address harms exacerbated by engagement-based recommender systems. For example, Europe's Digital Services Act requires platforms to provide at least one option for an alternative recommender system, or 'feed', that is not based on user profiling,[70] which is often a reverse chronological feed.[71]

However, research has found that users find reverse chronological feeds less engaging and often switch back to engagement-optimised feeds, despite voicing dissatisfaction with them.[72] These feeds may also mean users see less content from their social network.[73] Reverse chronological feeds are still vulnerable to manipulation, such as through content spamming,[74] which can increase users' exposure to abuse[75] and amplify the reach of untrustworthy content.[76]

There are several challenges in assessing reverse chronological feeds in a methodical and meaningful way. Platforms often do not provide details on how many users choose these feeds. They have also not been a significant focus of research to date because they are generally understood to be rarely used. Further, the limited research available tends to focus on how platform design influences political views, rather than harmful content more broadly.

What is clear is that there does **not need to be a binary choice** between optimising for engagement and using chronological feeds. Research into alternatives to engagement-based and reverse-chronological sorting suggest that **less harmful forms of algorithmic sorting are possible** and that platforms should avoid treating all engagement signals in the same way.[77]

---

[70] M Marsh, E L Podgorsek, Algorithm Watch, *A guide to the Digital Services Act, the EU's law to reign in Bic Tech,* 2026, https://algorithmwatch.org/en/dsa-explained/; TikTok, *Making your feed For You*, 2026, https://www.tiktok.com/safety/en/making-your-feed-for-you; YouTube, *Manage your recommendations & search results,* 2026, https://support.google.com/youtube/answer/6342839?hl=en.

[71] Meta, *User control: How to shape your experience on Meta's Platforms,* 2026, https://transparency.meta.com/features/user-control-shape-your-experience-on-meta-platforms; G de Seta, P K Friedman, Tech Policy Press, *Algorithms don't make the rules,* 2025, https://www.techpolicy.press/algorithms-dont-make-the-rules.

[72] A M Guess et al., 'How Do Social Media Feed Algorithms Affect Attitudes and Behavior in an Election Campaign?', *Science,* 2023, https://www.science.org/doi/10.1126/science.abp9364.

[73] D Paresh, Wired, *Meta jus proved people hate chronological feeds,* 2023, https://www.wired.com/story/meta-just-proved-people-hate-chronological-feeds; Moehring et al., 'Better Feeds: Algorithms that put people first'.

[74] Moehring et al., 'Better Feeds: Algorithms that put people first'.

[75] Moehring et al., 'Better Feeds: Algorithms that put people first'.

[76] Moehring et al., 'Better Feeds: Algorithms that put people first'.

[77] Moehring et al., 'Better Feeds: Algorithms that put people first'.

Recommender systems can be optimised to support **users' deliberate and long-term preferences,** rather than short-term engagement.[78] Giving greater weight to these preferences can provide users with more control. For example, clicking 'show less' on a piece of content, or answering surveys about the content a user sees can meaningfully shape recommendations over time, instead of relying on passive indicators of engagement like dwell time.

While not all users engage with these tools, platforms can build predictive models from the responses they do receive[79] and from other signals, such as followed accounts.[80] Currently, these tools are rarely used by platforms to promote long-term user value.[81] However, they can serve an important purpose when balanced with other factors, such as bridging across groups to avoid reinforcing in-group bias,[82] slowing the spread of viral content[83] and robust content moderation.

Understanding and designing safety tools and interventions is complex. Some changes may lead to **unexpected or unwanted user behaviour**, such as users engaging with safety features less or turning them off altogether.[84] For example, research found that providing explanations about how content is restricted through content controls made young users less likely to use these features. This may be because the explanation reminded participants that 'restricted' content exists, making them more curious about what they were missing.[85]

---

[78] Moehring et al., 'Better Feeds: Algorithms that put people first'.
[79] Milli et al., 'Engagement, user satisfaction, and the amplification of divisive content on social media'.
[80] Moehring et al., 'Better Feeds: Algorithms that put people first'.
[81] Milli et al., 'Engagement, user satisfaction, and the amplification of divisive content on social media'; A Narayanan, *Understanding social media recommendation algorithms*, 2023, https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms.
[82] Content bridging is a strategy designed to foster connection and community recommending content to users that differs from their current recommendations but is not so different that it is likely to cause a negative reaction. The goal of bridging is to increase mutual understanding and trust across divides, creating space for productive conflict, deliberation, or cooperation. For a more detailed discussion, see: Milli et al., 'Engagement, user satisfaction, and the amplification of divisive content on social media' and A Ovadya and L Thorburn, 'Bridging systems: Open problems for countering destructive divisiveness across ranking, recommenders, and governance', 2023, https://knightcolumbia.org/content/bridging-systems.
[83] J Garland et al., 'The case against efficiency: friction in social media', *npj Complex*, 2026, 3(5), https://www.nature.com/articles/s44260-025-00061-z.
[84] Ofcom, 'User empowerment and content control tools: Testing how design affects people's choices'.
[85] Ofcom, 'Behavioural insights to engage children with content controls: Testing online safety measures with children (13-17 years)', 2023**,** https://www.ofcom.org.uk/online-safety/safety-technology/boosting-childrens-safety-online-user-controls.

Education, literacy and user empowerment tools

Digital literacy is the foundation for people to interact online safely, competently and critically. It is important to **enhance digital literacy** and give all users the skills and confidence to manage their online experiences. This includes helping users understand how recommender systems shape the content they see and the online experience they have. Platforms and services can raise awareness of how their recommender systems work. This can contribute to **building algorithmic literacy**, as part of a broader effort towards building digital literacy.

Services should **provide users with empowerment tools**, including options to filter and limit certain content. For example, by limiting posts that mention certain terms. Other tools include settings that protect accounts from direct messages from strangers or from users they are not connected with on the platform. How services promote these tools is also important. If service providers are transparent and explain what happens when users apply content controls, the information is more likely to be useful and encourage users to engage with the safety tools available to them.[86]

Principles for designing successful user controls

Research by the Knight Georgetown Institute has outlined general design principles for successful user controls. These include:

- Providing both granular controls over individual pieces of content and broader controls over specific topics.
- Making controls that are transparent and easy to find.
- Placing controls prominently within users' settings to minimise the need for navigation.

These design principles still need to be tailored to each platform or service.

User empowerment tools should not be the only mechanism service providers rely on. While these tools can give users more control over their online experiences and help them engage with content more critically, they are not a substitute for safer platform and service design. User empowerment tools need to be backed by meaningful proactive measures so that the burden of safety does not fall solely on users. This is part of a comprehensive Safety by Design approach.

[86] Ofcom, 'User empowerment and content control tools: Testing how design affects people's choices'.

Designing safer systems

To address the ongoing concern about the role of algorithms and recommender systems in a range of online harms, made worse by rapidly shifting platform norms and new technologies, greater, concerted action is required.

Industry should take a proactive and deliberate approach to safer design. This includes giving users more control over algorithm settings and improving transparency about how recommender systems and algorithms are designed and used. Industry should make user safety the first and foremost priority. This includes the best interests of children and young people, rather than allowing those interests to be displaced by other considerations, including commercial factors.

It is important that approaches are layered, multi-faceted and holistic. Recommender systems amplify harms that already exist in online and offline settings. Measures that address recommender systems in isolation, without considering wider platform design and broader social contexts, may have limited impact.

In addition to urging industry to implement Safety by Design, we also encourage other key stakeholders in the online safety ecosystem, such as regulators and policymakers, to adopt the same approach.

Safety By Design

eSafety's **Safety by Design** initiative encourages technology companies to take proactive steps to embed user safety and invest in risk mitigation from the outset and throughout the product life cycle.

Platforms and service providers have a critical opportunity to embed safer design practices to minimise the risk from recommender systems by following the three foundational principles of Safety by Design: service provider responsibility, user empowerment and autonomy, and transparency and accountability.

Service provider responsibility

The burden of safety should never fall solely on users. Although recommender systems can be complex, users should not be expected to understand how they work to avoid harmful content, or exposure to it at a harmful scale.

Platform and service providers should mitigate harm by assessing the potential safety risks associated with their recommender systems and take active steps to address these

vulnerabilities before harm occurs. Platforms and services should continue testing the impact of their recommender systems and algorithms, including any changes made to them, on an ongoing basis.[87]

To support service provider responsibility, platform and service providers should implement measures such as:

- **optimising the algorithms that underpin recommender systems for quality-focused metrics, such as the authoritativeness or diversity of content**[88], instead of, or in addition to, engagement
- **establishing** and **enforcing content policies** and building them into the design of recommender systems
- actively **moderating harmful content and removing illegal content** that clearly violates relevant laws and platform terms of service or community standards
- introducing **human review as a circuit breaker** for content at risk of being amplified
- introducing **additional friction through design features**, such as content labels and prompts.

User empowerment and autonomy

Platform and service providers should ensure their products and platforms uphold the dignity of users, align with their best interests and uphold fundamental human rights. This involves an understanding that online harms can often be intersectional. The structures and features of platforms and services, including the design of recommender systems and algorithms, can reinforce existing societal inequalities. It is essential to consult a diverse group of individuals and communities, including marginalised and underrepresented individuals and communities, throughout all stages of platform or service design, development and implementation.

To support user empowerment and autonomy, platform and service providers should implement measures such as:

- **opt-out settings** that give users choice, ownership and control over the types of content they receive, and the ability to opt out of others
- **greater choice for users**, including **controls and alternative curation models** for their feeds, as well as **clear feedback loops** for content moderation decisions

[87] Moehring et al., 'Better Feeds: Algorithms that put people first'.
[88] This should be subject to consultation, public scrutiny and testing in determining what sources are authoritative.

- **behavioural cues and prompts** that help users establish and reinforce positive patterns of behaviour
- **educative prompts or nudges** that support users to develop their algorithmic literacy.[89]

Transparency and accountability

Platform and service providers should share information with users, regulators and researchers about how their recommender systems operate. They should also make information available about design choices, the objectives of the algorithms used by their service and the outcomes of their recommender systems. This can help users make informed choices about how they interact with information on their feeds.

To support transparency and accountability, platform and service providers should implement measures such as:

- providing **public information about how users respond to prompts based on engagement or interactions with recommended content**, such as the nature of those responses, including whether they are positive or negative
- enhancing **transparency reporting and auditing practices**, including by making more information publicly available to researchers, experts and regulators. This helps to improve future interventions and regulation, and builds trust with users
- engaging in **international multi-stakeholder bodies that develop technical standards** for the ethical use and deployment of recommender systems, to support a shared understanding of best practice.

Reform options

The Australian Government is progressing reforms to strengthen the *Online Safety Act 2021* (Cth) following an independent review conducted by Ms Delia Rickard PSM. The Government published its **response to this review** on 14 April 2026.

The review affirmed eSafety's role and the vital work we do to protect Australians from online harm. It also noted harms associated with recommender systems and algorithms, including promoting harmful content, amplifying extremist views and recommending unsafe accounts. It suggested these risks could be addressed under a duty of care model.

[89] Algorithmic literacy is a component of digital literacy. Digital literacy encompasses the knowledge and skills individuals need to create, manage, communicate and investigate data, information and ideas, and solve problems when engaging with technologies and in online environments.

The Government has committed to legislate a digital duty of care, which will require platforms to exercise due diligence and take steps to prevent foreseeable harms on their services. In total, the Government intends to implement or further consider 64 of the review's 67 recommendations.

We look forward to continuing to work with the government to develop and implement a duty of care that creates robust and enforceable obligations. Building on eSafety's work to date to put the onus on services to make their platforms safer, through both our Safety by Design initiative and systemic regulatory schemes, we will work towards a duty of care that addresses online harms comprehensively and proactively, including those relating to recommender systems and algorithms.

Reforms in online safety will work in conjunction with reforms in other areas, including privacy and artificial intelligence. This will support a whole-of-government, cross-portfolio approach to addressing the harms and promoting the benefits of recommender systems and algorithms.

Looking ahead

Technology itself is rarely inherently harmful. Rather, harms arise from the ways in which otherwise neutral technologies are designed and deployed.

Many recommender systems are not designed or deployed in ways that prioritise user safety. They are created to maximise engagement and keep users on the platform. This results in the balance between their benefits and risks often being out of step. Without intervention, these risks will continue to increase, amplify and compound – negatively impacting both individual users and broader societal safety and wellbeing.

While there is growing momentum to urgently address these harms in a meaningful way, improving recommender systems is a complex challenge. Recommender systems are only one factor in how online harms emerge, meaning mitigation strategies must navigate multiple, often competing considerations. While it is important to acknowledge this complexity, it must not be used to deter or dilute necessary action.

There is a clear need to both incentivise and compel industry to strengthen its approach to recommender systems, particularly in relation to transparency, accountability, and ensuring that users' best interests are central to platform design.

A holistic approach to safer design is essential. Recommender systems are best understood and addressed as part of a whole systems safety approach to the wider online ecosystem, rather than as the sole driver of online harms. They should therefore be addressed through a layered and multidimensional strategy to improve online safety.

Going forward, we will continue to draw on our existing regulatory levers to address recommender systems, prioritising education, outreach and awareness raising, while working towards stronger systemic reform under a duty of care.

Further information and resources

If you or someone in your care is experiencing serious online abuse or harm, you can **make a report to eSafety**. You can also speak to a mental health professional through an **expert counselling and support service**.

Report incidents of child sexual exploitation and abuse material to the **Australian Centre to Counter Child Exploitation**. If you believe a child is in immediate danger, call 000 or contact your local police station.

eSafety is developing education and training programs to raise awareness of the impacts of recommender systems, including both their opportunities and risks, and the tools available to manage them.

Targeted online safety education includes:

- **webinars** for educators, youth serving professionals, and parents and carers about recommender systems and algorithms, including **how algorithms can influence and reinforce harmful beliefs**, and **the rewards and risks of recommender systems for young people**
- discussion through **key advisory and consultation mechanisms**, including with the **Trusted eSafety Providers**, **National Online Safety Education Council** (which includes ACARA – the Australian Curriculum and Assessment Authority), eSafety's Parent Advisory Group, and the **eSafety Youth Council**, to inform the development of information on **Drifting into an 'echo chamber'? Take control of the algorithms that shape your feed**
- published **advisories on topics** such as **how algorithms are shaping our adolescents** and **how violent content is reaching children and what you can do**
- updates to eSafety's **Toolkit for Schools** and **Best Practice Framework for Online Safety Education** to include guidance on a whole-of-school approach to online safety education and support prevention and responses to online harms. Additional resources to **support digital literacy** are also available.