Ending artificial amplification of hate & hysteria

Rapidly resolving the recommender system crisis

Irish Council for Civil Liberties

In this note

Revision 1.1

Contact: Dr Johnny Ryan (johnny.ryan@iccl.ie).

Cover photo by Adrian Weckler.

Enforce

Enforce is a unit of the Irish Council for Civil Liberties (ICCL). Learn more about our work at https://www.iccl.ie/enforce/

Summary

Social media was supposed to connect us. Instead, it tears society apart. While many things contribute to public unrest, Big Tech's algorithmic "recommender systems" are the difference between a tiny group of angry people online and large riots that wreck our cities.

November's riot in Dublin is a shock that should stimulate action. If this can happen in Dublin, it can happen anywhere in Europe. We need an urgent solution before more cities explode.

European audiovisual media regulators should require digital platforms to switch off dangerous recommender systems for all non-age proven persons, particularly where those recommender systems process "special category" personal data that are particularly protected under the GDPR.

Digital platforms should not be allowed to build intimate profiles about our children – or any person whose age is unproven – in order to then manipulate them for profit by artificially amplifying hate, hysteria, and disinformation in their personalised feeds.

Audiovisual media regulators have the power to do so under Article 6a(1) of the AVMSD, which empowers them to protect minors against potential harms. The practical limits of age verification make it impossible to distinguish minors online. To protect minors, regulators must protect all non-age proven persons.

Coimisiún na Meán, Ireland's audiovisual media regulator, has issued a draft binding code for video platforms that requires them to stop automatically using recommender systems based on intimate profiles of each user. People - not Big Tech's algorithms - should be free to decide what they see and share. We believe that this approach should be applied in every Member State.

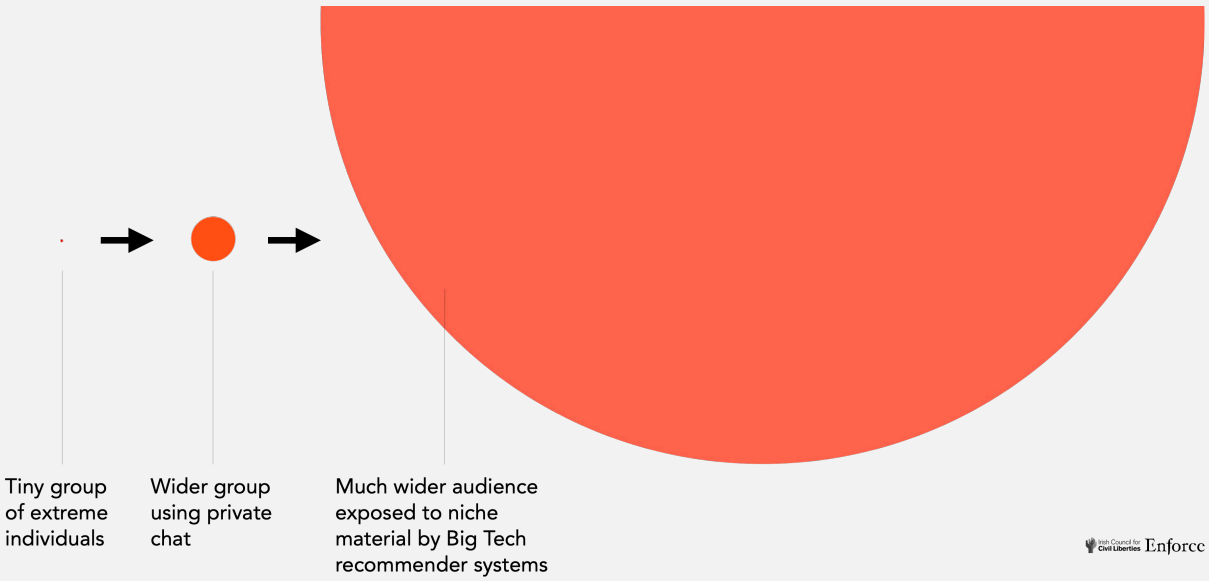
Problem: amplifying hate and hysteria

An algorithmic "recommender system" selects emotive and extreme content and shows it to people it estimates are most likely to be outraged. These people then spend longer on the platform, which allows the platform to make more money showing them ads.

- **Meta's own internal research** reported that "64% of all extremist group joins are due to our recommendation tools… **Our recommendation systems grow the problem**".[1]
- Nearly three quarters of the problematic[2] content seen by 37,000+ test volunteers on YouTube was due to the **YouTube's recommender system** amplifying it.[3]
- In August 2023 an Anti Defamation League study found that Facebook, Instagram, and X recommended antisemitic and conspiracy content to 14 year old test users.[4]
- The European Commission reports that Russian disinformation about Ukraine was achieved by pro-Kremlin actors and "algorithmic recommendation by the platforms".[5]
- U.N. investigators found that Meta played a "determining role" in **Myanmar's 2017 genocide**.[6] Amnesty International reported Meta's algorithms were key contributors.[7]
- Just one hour after Amnesty started a **TikTok** account posing as a **13 year old child** who views mental health content, videos encouraging **suicide** were recommended.[8]

From niche extremism to big problem

Recommender systems transform tiny extremist groups into large social problems.
Without algorithmic amplification, material from niche extremists are not widely seen.

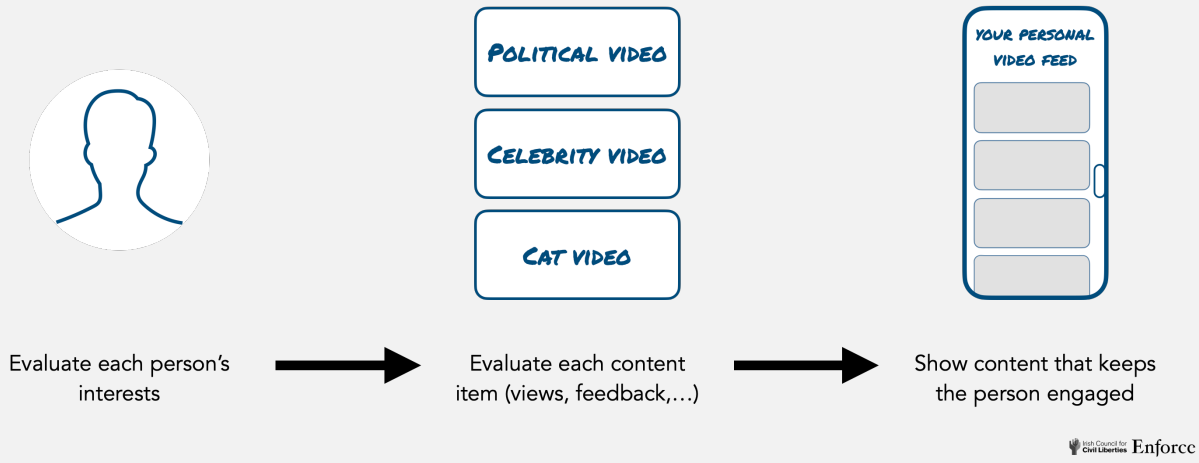


How recommender systems process data

Recommender systems process two main sets of data:

1. Data about each item of content uploaded to the platform, such as how many times each item is viewed; and
2. **Personal data** about each person who uses the platform, such as their viewing habits and interests.[9]

The system then selects items of content that it estimates will keep the person engaged (watching, commenting, sharing, etc.) the longest. The longer a person is engaged, the more they can be monetised by the platform. Since emotive posts that evoke hate and hysteria are highly engaging, they are amplified, and the platform monetises this.



Social media were supposed to be notice boards where people choose what they share with their friends. Instead, social media became toxic places where Big Tech feeds us a toxic diet of hate and hysteria. Social media was supposed to connect us. Instead, it is tearing society apart.

Big Tech's inadequate response

Digital platforms have a very poor record of self-improvement and responsible behaviour, even when lives are at stake as in Myanmar's genocide. Even when a platform understands the harm its recommender system causes, it is unlikely to voluntarily act. Despite internal concern about amplifying hazardous content, from 2017 to 2020 Meta strongly amplified[10] posts that received "emoji" reactions from other people. Then, despite internal research in 2019 confirming that content receiving "angry emojis" was more likely to be misinformation, it persisted until late 2020.[11]

Solution: off by default

Coimisiún na Meán, Ireland's audiovisual media regulator, has issued a draft binding code for video platforms that requires them to stop automatically using recommender systems based on intimate profiles of each user. We believe this approach should be applied by every Member State.

Digital platforms should not be allowed to build intimate profiles about our children – or any person whose age is unproven – in order to then manipulate them for profit by artificially amplifying hate, hysteria, and disinformation in their personalised feeds.

Recommender systems that use information about people's political and philosophical views should be off by default in order to comply with Article 9 of the GDPR and Article 6a(1) of the AVMSD. People – not algorithms – should decide what they see on digital platforms.

This is an efficient and rights-respecting approach. Switching off recommender systems by default prevents algorithmic amplification of hate, hysteria, and disinformation without requiring burdensome moderation of individual items of content. A 2019 internal Meta document (leaked by whistle-blower Frances Haugen) affirmed that content moderation is impossible, and the focus should instead be on stopping artificial amplification:

> *"We are never going to remove everything harmful from a communications medium used by so many, but we can at least … stop magnifying harmful content by giving it unnatural distribution".*[12]

Since identifying and moderating all dangerous content is not possible, curbing algorithmic amplification is likely to be far more effective. Crucially, this **avoids intrusion upon to the right to freedom of expression**: instead of limiting speech, it stops artificial amplification.

There are two tools available to rapidly achieve this.

First, **GDPR Article 9** applies to any recommender systems that rely on processing data that could reveal a person's political or philosophical views. GDPR Article 9 provides particular protections for "special categories of personal data":

> *"…personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation*
>
> *2. Paragraph 1 shall not apply if one of the following applies: (a) the data subject has given explicit consent to the processing of those personal data for one or more specified purposes, …"*

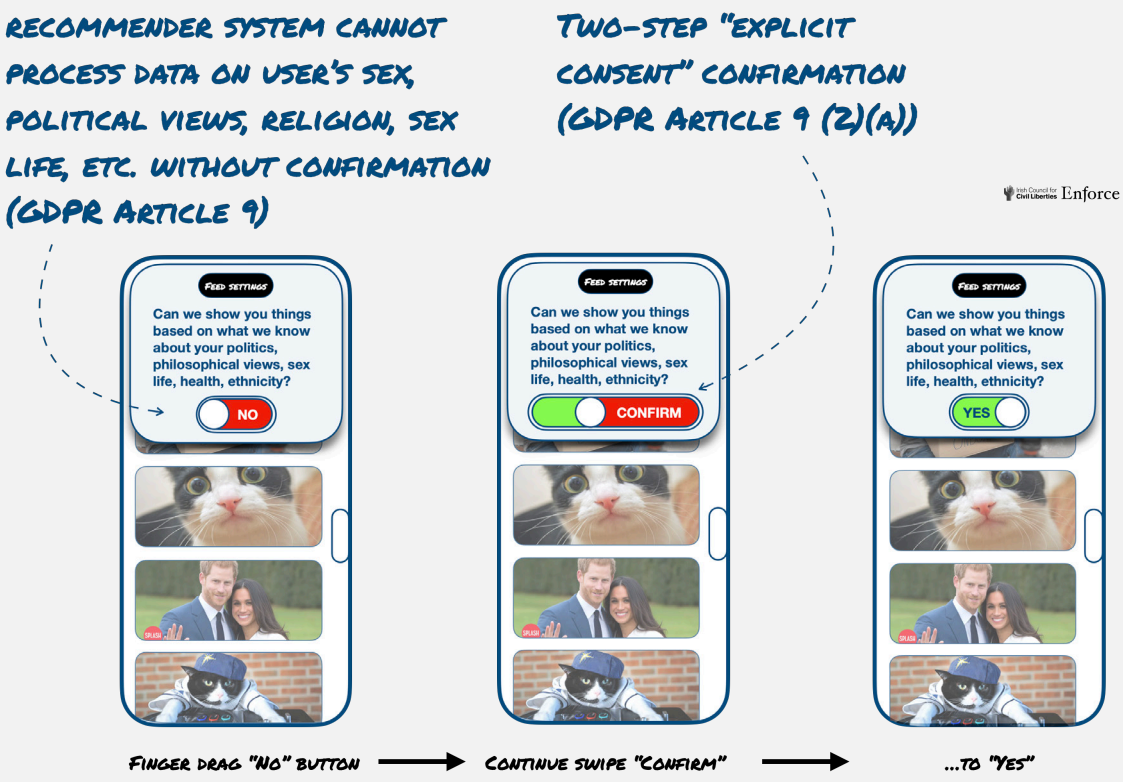
This is directly relevant to recommender systems that amplify hate, hysteria, and disinformation. A recommender system will process "special category" personal data about a person in order to decide what content to show to them. For example, the algorithm will identify whether or not each individual person is likely to be outraged by a video about immigration in order to decide whether to put that video in each person's feeds.

Such data allow special category personal data to be revealed. As the CJEU affirmed in July 2023, it does not matter whether the platform intends to obtain this information, nor whether the information is correct. Rather, the objective of Article 9 of the GDPR is to prohibit processing of special category personal data, irrespective of its stated purpose.[13]

The sole relevant derogation under Article 9 that would allow a recommender system to process special category personal data is if a person has given "explicit" (two-step)[14] consent. However, it does not appear that this two-step explicit consent has ever been sought by a platform for this reason. Thus, all recommender systems that process special category personal data do so unlawfully. They should not be operating at all.

Hate, hysteria, disinformation must be off by default

"Explicit consent" requires a two-step action to give a person the opportunity to confirm their consent.[15] This must occur before a platform processes any data that reveal a person's racial or ethnic origin, political opinions, religious beliefs and data concerning health or sex life or sexual orientation. An indicative design of this two-step consent is below.



Second, **AVMSD Article 6a(1)** applies to all, or almost all, recommender systems for persons who are not age proven (everyone, with extremely narrow exceptions).

The DSA and AVMSD are complementary, as Article 2(4) and Recital 68 of the DSA observe. AVMSD Article 28b(1) defines three categories of person who must be protected against potential or actual specified harms. The first category of person and potential harm, defined at point (a), is:

> *"…minors from programmes, user-generated videos and audiovisual commercial communications which may impair their physical, mental or moral development in accordance with Article 6a(1);"*

AVMSD Article 6a(1) (cited above in Article 28b(1)) provides that:

> *"Member States shall take appropriate measures to ensure that audiovisual media services provided by media service providers under their jurisdiction which may impair the physical, mental or moral development of minors are only made available in such a way as to ensure that minors will not normally hear or see them. Such measures may include selecting the time of the broadcast, age verification tools or other technical measures. They shall be proportionate to the potential harm of the programme.*
>
> *The most harmful content, such as gratuitous violence and pornography, shall be subject to the strictest measures."*

The criteria set by the legislator has two components. First, "may impair" or "potentially harm". Second, a possible harm to the "physical, mental or moral development of minors".

The regulator must act to ensure that recommender systems "are only made available in such a way as to ensure that minors will not normally hear or see them". But there are no technical means of definitively confirming whether a person using a platform is a minor or an adult (with the narrow exception of certified passport or national identity cards). Existing "age verification" methods are unreliable, circumventable, and legally fraught. (See recent developments in Australian legislation,[16] and separate reports from CNIL, the French data protection authority,[17] and UK Ofcom.[18])

The known limitations of age verification mean that Article 6a(1) of the AVMSD empowers regulators for audiovisual media services to protect all non-age proven persons from "normally hear[ing] or see[ing]" potentially harmful content. Indeed, this is the only way to realise the objective of Article 6a(1).

Audiovisual regulatory authorities can therefore act against the acute potential and real harms of dangerous recommender systems by directing platforms to switch them off by default. Article 6a(1) provides for proportionality, so that lower severity content intended for adults, such as dating websites etc., are not unduly impacted.

What happens when recommender systems switch off?

Algorithmic recommender systems are not technically essential components of digital platforms. Virtually all websites and news media operate without such systems, relying instead on the curatorial art of their editors. Indeed, Article 38 of the Digital Services Act provides that recommender systems based on a profile must be optional.

Nor are these systems legally essential. The European Court of Justice (CJEU) ruled in July 2023 in *Bundeskartellamt v Meta* that personalisation of content is "not objectively indispensable".[19]

There are alternative methods to curate a digital platform and show users a mix of memes, cat videos, celebrity news, and unboxing videos that do not require recommender systems which process profiles of each user.

For example, platforms may rely on one or more of the following – and currently do so in addition to their recommender systems or as part of them:

- the user's selection from a menu of the categories of content they are interested in;
- expert editors curating categories of video and video creators;
- ranking content by other factors, such as number of views, reputation of author/ producer, quality rating feedback users, etc.

Despite the power of alternatives, some platforms may respond with "**malicious compliance**" by implementing poor experiences, in order to provoke outcry against regulation. The following, for example, must not be accepted from the platforms: switching from a recommender system to an entirely unedited and unordered feed of randomised video, and then prompting users to switch back on the recommender system. Moreover, digital platforms who maliciously comply create the risk that their users will depart to competitors who offer better service. Malicious compliance should be damaging to the platform, not the user.

Notes

[1] "Facebook Executives Shut Down Efforts to Make the Site Less Divisive", Wall St. Journal, 26 May 2020 (URL: https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499). This internal research in 2016 was confirmed again in 2019.

[2] "YouTube Regrets: A crowdsourced investigation into YouTube's recommendation algorithm", Mozilla, July 2021 (URL: https://assets.mofoprod.net/network/documents/Mozilla_YouTube_Regrets_Report.pdf), pp 9-13.

[3] ibid. p. 17.

[4] "From Bad To Worse: Amplification and Auto-Generation of Hate", ADL, 16 August 2023 (URL: https://www.adl.org/resources/report/bad-worse-amplification-and-auto-generation-hate)

[5] "Digital Services Act: Application of the Risk Management Framework to Russian disinformation campaigns", European Commission, 30 August 2023 (URL: https://op.europa.eu/en/publication-detail/-/publication/c1d645d0-42f5-11ee-a8b8-01aa75ed71a1/language-en), p. 64.

[6] U.N. investigators cite Facebook role in Myanmar crisis, Reuters, 12 March 2018 (URL: https://www.reuters.com/article/us-myanmar-rohingya-facebook/u-n-investigators-cite-facebook-role-in-myanmar-crisis-idUSKCN1GO2PN).

[7] "The social atrocity: Meta and the right to remedy for the Rohingya", Amnesty International, 2022 (URL: https://www.amnesty.org/en/documents/ASA16/5933/2022/en/), pp. 45-48, p. 71.

[8] https://www.amnesty.org/en/latest/news/2023/11/tiktok-risks-pushing-children-towards-harmful-content/.

[9] Davidson et al., "The YouTube Video Recommendation System", RecSys '10: Proceedings of the fourth ACM conference on Recommender systems, September 2010 https://bytes.usc.edu/cs572/s23-sear-chhh/lectures/YouTube/docs/The_YouTube_video_recommendation_system.pdf; see also https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/

[10] Five times the amplification of a standard "like".

[11] "Five points for anger, one for a 'like': How Facebook's formula fostered rage and misinformation", *Washington Post*, 26 October 2021 (URL: https://www.washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/).

[12] The Facebook Papers, "We are Responsible for Viral Content", 11 December 2019, p.17

[13] Paragraphs 68-70 of CJEU decision in case Case C-252/21 (URL: https://curia.europa.eu/juris/document/document.jsf?text=&docid=275125&pageIndex=0&doclang=en&mode=lst&dir=&occ=first&part=1&cid=2705560).

[14] "Guidelines on Consent under Regulation 2016/679 (wp259rev.01)", Article 29 Working Party (URL: https://ec.europa.eu/newsroom/article29/items/623051/en).

[15] "Guidelines 05/2020 on consent under Regulation 2016/679", European Data Protection Board, 4 May 2020 (URL: https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_202005_consent_en.pdf), pp. 20-22.

[16] In August 2023 the Australian Parliament concluded that it could not lawfully legislate for the age verification requested by the Australian e-Safety Commissioner. See https://www.theguardian.com/australia-news/2023/aug/31/roadmap-for-age-verification-online-pornographic-material-adult-websites-australia-law.

[17] CNIL, the French data protection authority, reported in 2022 that age verification is "circumventable and intrusive".https://www.cnil.fr/en/online-age-verification-balancing-privacy-and-protection-minors

[18] Ofcom's 2022 study of online user ages demonstrates the difficulty of achieving certainty of a person's age online. https://www.ofcom.org.uk/__data/assets/pdf_file/0015/245004/children-user-ages-chart-pack.pdf

[19] CJEU judgement of 4 July 2023, Bundeskartellamt v Meta, C-252/21, ECLI:EU:C:2023:537, paragraph 102.