A binding Code against toxic algorithms

ICCL submission to the Media Commission call for input on video-sharing platform services

SEPTEMBER 2023

Dr Johnny Ryan FRHistS
Senior Fellow, ICCL

Irish Council for
Civil Liberties

In this submission

Contact: johnny.ryan@iccl.ie

About ICCL

The Irish Council for Civil Liberties (ICCL) is Ireland's oldest independent human rights body. It has been at the forefront of every major rights advance in Irish society for almost half a century. ICCL helped legalise homosexuality, divorce, and contraception. We drove police reform, defending suspects' rights during dark times. In recent years, we led successful campaigns for marriage equality and reproductive rights.

Dr Johnny Ryan FRHistS is a Senior Fellow at ICCL. Previously he served in senior roles in technology and media. He is regularly invited to give expert testimony and has appeared before the European institutions and the U.S. Senate. His expert commentary has appeared in *The Economist*, *NATO Review*, and *The New York Times*.

Thanks to **Olga Cronin** and **Katarzyna Szymielewicz**.

Summary: act on algorithms

This submission demonstrates the hazard of platforms' algorithmic recommender systems, and proposes verifiable measures.

Selected Media Commission questions:

- Question 1 – "What do you think our main priorities and objectives should be in the first binding Online Safety Code for VSPS? What are the main online harms you would like to see it address and why?"

- Question 4 – "What approach do you think we should take to the level of detail in the Code? What role could non-binding guidance play in supplementing the Code?"

- Question 20 – "What approach do you think we should take in the Code to address feeds which cause harm because of the aggregate impact of the content they provide access to? Are there current practices which you consider to be best practice in this regard?"

Summary:

- Our submission focuses on **digital platforms' algorithmic amplification of hazardous content such as incitement to hate, violence and terrorism, racism and xenophobia**.

- We respond to questions 1, 4, and 20 of the Media Commission's invitation. Our answer to question 1 is the section "Recommender systems"; question 4 is the section "Prescriptive and verifiable"; and question 20 is the section "Action on algorithms".

- The section "Recommender systems" shows that **platforms' recommender systems are particularly dangerous**. The section "Prescriptive and verifiable" shows that **platforms' voluntary and discretionary measures are ineffective**.

- We suggest several measures. Primary among them is that the Code should mandate that algorithmic recommender systems are not activated by default by platforms. **Toxic algorithms must stay off until a user decides to switch them on.** People must be able to use digital platforms without algorithms injecting poison into their feeds.

- Acting against algorithmic amplification rather than attempting to identify and unpublish harmful content is likely to be more effective, and **avoids intrusion upon the right to freedom of expression**.

Recommender systems

RESPONSE TO MEDIA COMMISSION QUESTION 1

Recommender systems are understood to be dangerous, and require prioritisation.

Examples:

- In August 2023 an Anti Defamation League study found that **Facebook**, **Instagram**, and **X** (**Twitter**) recommended **antisemitic and conspiracy content** to test users, including to users as young as **14 years old**.[1]

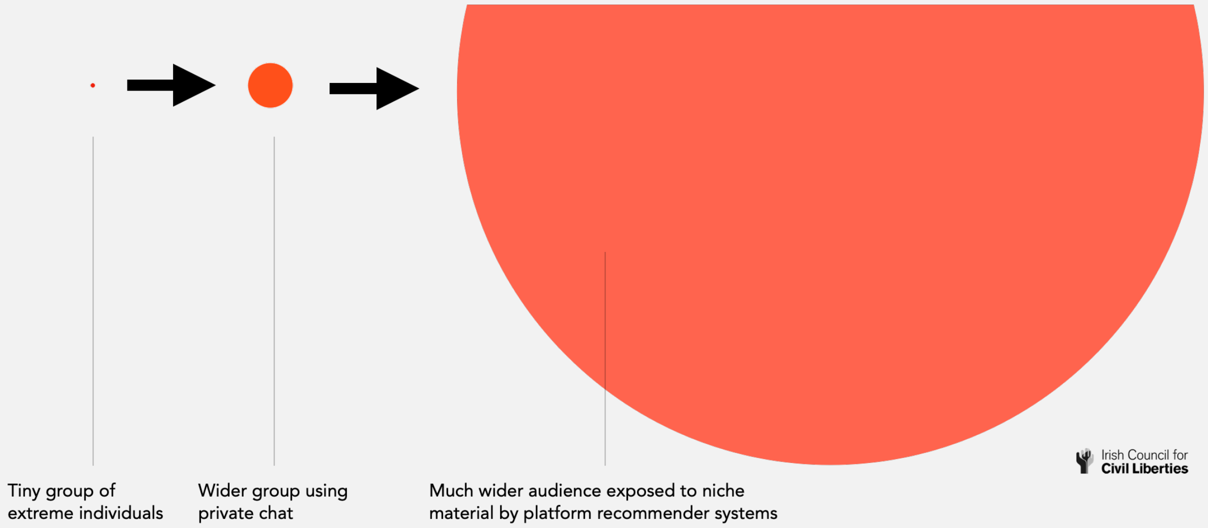
- A global study of 37,000+ YouTube volunteers in 2022 showed that **most (71%) of the problematic[2] content they saw on YouTube was presented to them by YouTube's recommender system**.[3] This new research followed YouTube recommender scandals and purported fixes by the company in preceding years.[4]

- In 2016 internal **Meta** research (later disclosed by whistleblower Frances Haugen) concluded that:

> "**64% of all extremist group joins are due to our recommendation tools**… Our recommendation systems grow the problem".[5] The researchers concluded: "Our algorithms exploit the human brain's attraction to divisiveness."[6]

Amplification of hate and hysteria

Digital platform **recommender systems** find emotive videos and posts and expose them to large audiences to maximise engagement. **Without algorithmic amplification, dangerous material from the small core group would not be widely seen**.


Tiny group of extreme individuals
Wider group using private chat
Much wider audience exposed to niche material by platform recommender systems
Irish Council for Civil Liberties

- An internal Meta document dated 2019 discussed "hate speech, divisive political speech, and misinformation" and noted:

> "compelling evidence that our core product mechanics, such as virality, **recommendations, and optimizing for engagement, are a significant part of why these types of speech flourish on the platform. … The mechanics of our platform are not neutral**".[7]

- Another 2019 internal Meta document concluded that content moderation is impossible at large scale, and the focus should be on avoiding algorithmic amplification of the content:

> "**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".[8]

- **United Nations investigators reported that Meta (Facebook) had played a "determining role" in Myanmar's 2017 genocide**.[9] **Amnesty International's** follow-on investigation reported that Meta's **algorithms** were essential contributors. Amnesty concluded that "**content-based solutions will never be sufficient to prevent and mitigate algorithmic harms**".[10]

- The **European Commission** reports that **Russian disinformation about its invasion of Ukraine** "was achieved through a combination of direct action by pro-Kremlin actors and **through algorithmic recommendation by the platforms**".[11]

Recommendation:

- Recommender systems find emotive content and expose it to large audiences to maximise engagement. **Without this algorithmic amplification, dangerous material from a tiny number of extremists would not be widely seen.**

- As the examples above show, the content covered by section 139K(2)(c) OSMR is far broader than the illustrative examples in point 5.3.5 of the Media Commission's request for input on recommender systems. **Since at least as early as 2016, digital platforms have understood that their recommender systems amplify hate and hysteria**.

- The Media Commission should therefore **prioritise acting against hazardous recommender systems** over other actions to tackle incitement to hate and violence, racism and xenophobia, and incitement to terrorism.*

* This recommendation does not relate to harms such as bullying, self-harm, child sexual abuse, etc. Other measures, such as content moderation and tackling addictive design will be required for other harms.

- Acting against algorithmic amplification rather than attempting to identify and unpublish harmful content is likely to be more effective, and **avoids intrusion upon the right to freedom of expression**.

Prescriptive and verifiable

RESPONSE TO MEDIA COMMISSION QUESTION 4

Voluntary and discretionary measures by platforms will not be sufficient.

Key insights:

- **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[12] 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 in strongly amplifying them until late 2020.[13]
- Digital platforms' voluntary measures against the risk they create are inadequate. In August 2023, the **European Commission** reported that voluntary measures taken by **YouTube**, **Facebook**, **Instagram**, **TikTok**, **Twitter**, and **Telegram** against Russian disinformation on their platforms had "failed".[14] It concluded that **"Article 35 [DSA] standards of effective risk mitigation were not met in the case of Kremlin disinformation campaigns"**.

Recommendation:

- The Code must be **binding**. It must be robustly enforced, if necessary, by application for a blocking order to the High Court.
- Measures required by the Code **must be practical to monitor**. Our recommendations in response to question 20 are designed with this in mind.
- Digital platforms should have no opportunity to evade their responsibilities. **Clarity is essential** in the Code's specification of mandatory measures.

Action on algorithms

RESPONSE TO MEDIA COMMISSION QUESTION 20

Algorithmic recommender systems are optional - and highly hazardous - features rather than intrinsic elements of digital platforms.

Key insights:

- Section 139K(4)(a) OSMR provides that a Code may provide for "standards that services must meet, practices that service providers must follow, or **measures that service providers must take**". The Media Commission is empowered to enforce those standards, including by way of an application to the High Court for a "blocking order" under section 139ZZC OSMR.

- Algorithmic recommender systems are neither legally nor technically essential components of digital platforms. The **European Court of Justice** (CJEU) ruled in July 2023 in *Bundeskartellamt v Meta* (including Facebook and Instagram) that **personalisation of content is "not objectively indispensable"**.[15] In addition, platforms are required by Article 38 DSA to provide **alternative recommendations not based on a profile of the user**.

- Switching algorithmic recommender systems off is technically trivial. Virtually all websites and news media operate without such systems, **relying instead on the curatorial art of their editors**.

- 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 the user's selection from a menu of the categories of content they are interested in, and have expert editors curate those categories of video and video creators.

- Digital platforms are required by **Article 9 GDPR** to have the person's "explicit consent" to process "special category" personal data, including inferences about the platform user's political views, sexuality, religion, ethnicity, health. These data cannot be processed for a recommender system unless the person has given their consent. **Any recommender systems that engage with a user's politics, sexuality, religion, ethnicity, or health must be off by default.**

Recommendations:

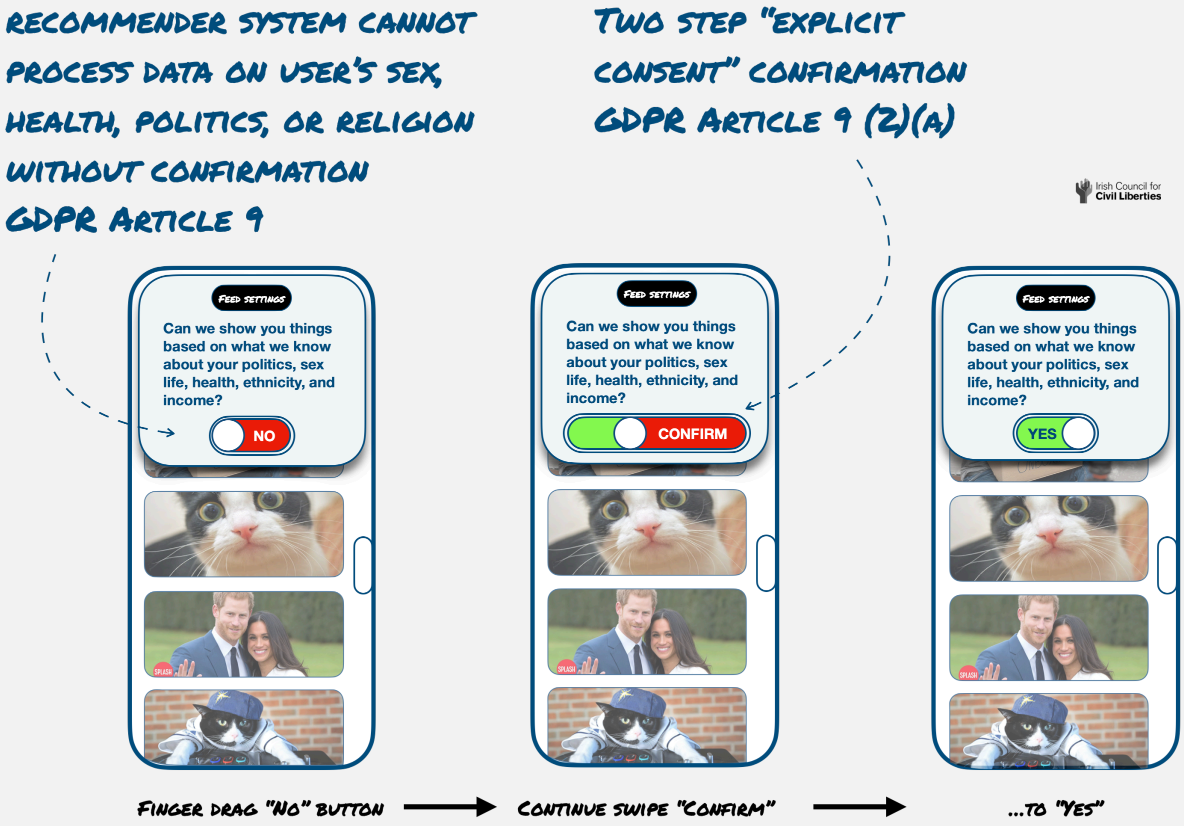
- The Code should mandate that algorithmic recommender systems are **not activated by default** by platforms. Users must be able to use a platform without being exposed to toxic algorithms that inject poison into their feeds.

- This should apply generally, but in particular to recommender systems that process (including by inference or proxy) "special category" data as defined by Article 9 GDPR. The GDPR prohibits processing of data about people's **health, sexuality, political and philosophical views, religious beliefs and ethnicity**. The only applicable derogation for a platform is if a user has given "explicit consent".

- The Code should require platforms to implement **lawful requests for explicit consent**.

Politics, sexuality, health... off by default

"Explicit consent" is understood to require a two-step action to give the person the opportunity to confirm their consent.[16] Our indicative design two-step action is below.


RECOMMENDER SYSTEM CANNOT PROCESS DATA ON USER'S SEX, HEALTH, POLITICS, OR RELIGION WITHOUT CONFIRMATION GDPR ARTICLE 9
TWO STEP "EXPLICIT CONSENT" CONFIRMATION GDPR ARTICLE 9 (2)(A)
Irish Council for Civil Liberties
FEED SETTINGS
NO
CONFIRM
YES
FINGER DRAG "NO" BUTTON
CONTINUE SWIPE "CONFIRM"
...TO "YES"

Can we show you things based on what we know about your politics, sex life, health, ethnicity, and income?

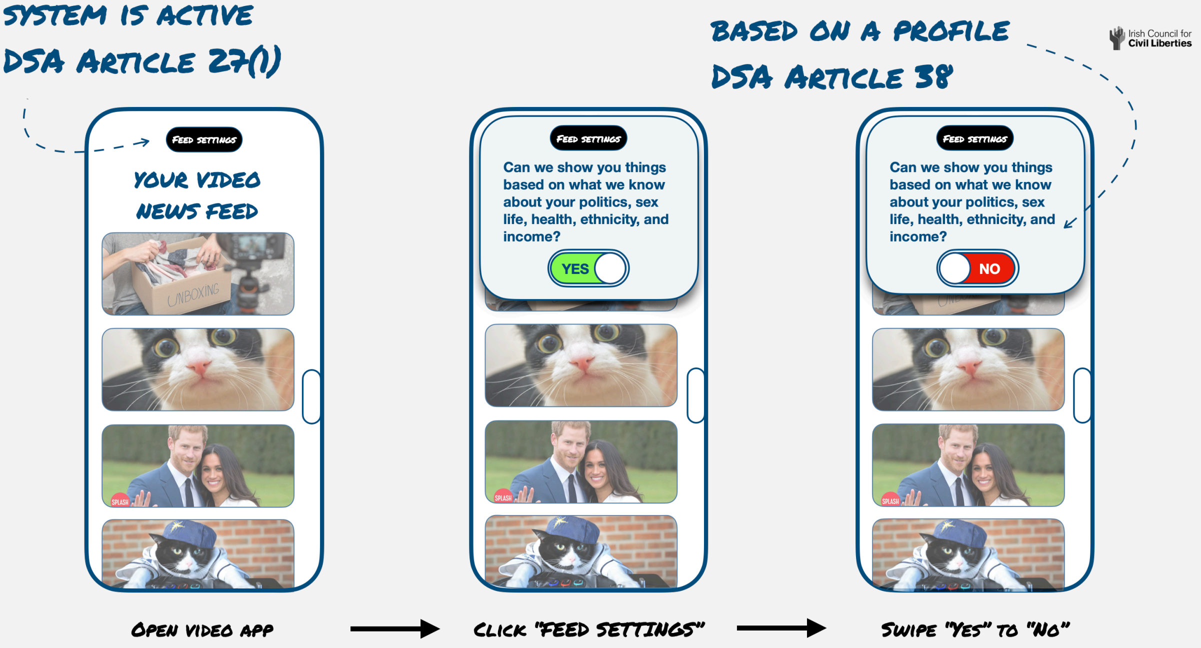
- The Code should require that **if a user activates a recommender system, then an immediately visible means of deactivating that recommendation system is shown prominently on the screen at all times** where the system is active, as provided for in DSA Article 27(1) and Article 38 of the DSA.

The DSA recommender system "off" switch

The Digital Services Act requires digital platforms to provide a recommender system off-switch, which must be visible at all times when the recommender system is active. Our indicative design for this is below.

OPTIONS MUST BE VISIBLE WHERE THE RECOMMENDER SYSTEM IS ACTIVE DSA ARTICLE 27(1)

MANDATORY OPTION FOR A RECOMMENDER SYSTEM NOT BASED ON A PROFILE DSA ARTICLE 38

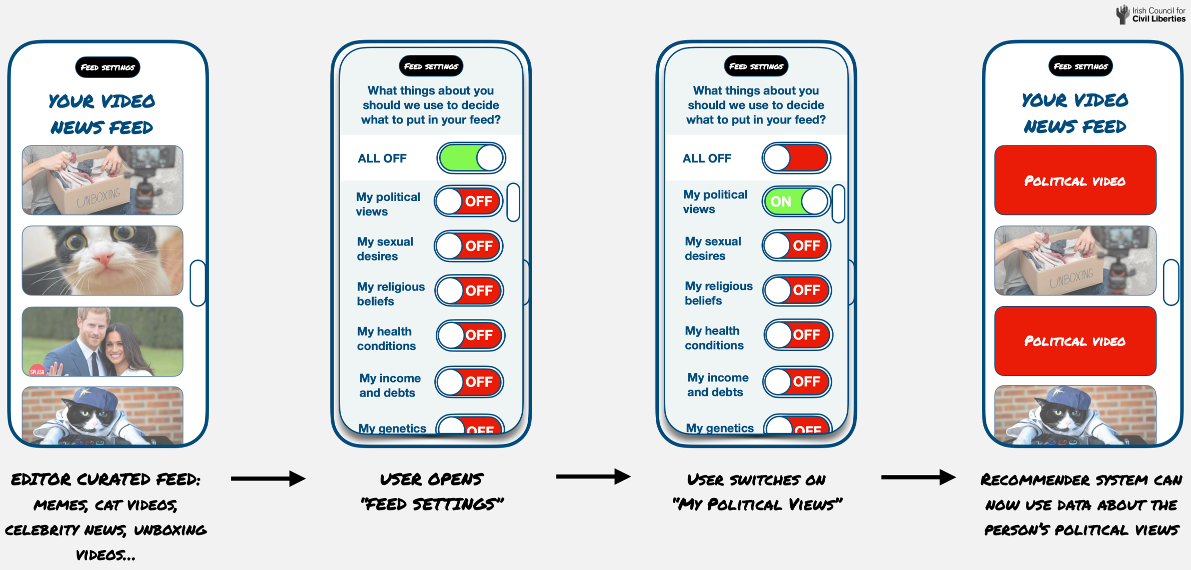


Can we show you things based on what we know about your politics, sex life, health, ethnicity, and income?

- The Media Commission may wish to consider whether the Code should also mandate granular user control over the activation of recommender systems, including the types of data about the user available to a recommender system.

Granular control

A user may wish to receive algorithmic recommendations related to their financial situation without the recommender system also making inferences about other intimate aspects of their character and circumstances. Our indicative design for granular control is below.



What things about you should we use to decide what to put in your feed?

ALL OFF ON

My political views OFF

My sexual desires OFF

My religious beliefs OFF

My health conditions OFF

My income and debts OFF

My genetics OFF

What things about you should we use to decide what to put in your feed?

ALL OFF OFF

My political views ON

My sexual desires OFF

My religious beliefs OFF

My health conditions OFF

My income and debts OFF

My genetics OFF

EDITOR CURATED FEED: MEMES, CAT VIDEOS, CELEBRITY NEWS, UNBOXING VIDEOS... → USER OPENS "FEED SETTINGS" → USER SWITCHES ON "MY POLITICAL VIEWS" → RECOMMENDER SYSTEM CAN NOW USE DATA ABOUT THE PERSON'S POLITICAL VIEWS

- The Media Commission should be prepared for the possibility that platforms will respond with "**malicious compliance**": implementing the least attractive designs and experiences for users in order to provoke outcry against regulatory intervention. For example, an entirely unedited and unordered feed of randomised video. However, digital platforms who maliciously comply create the risk that their users will depart to competitors who offer better service. Malicious compliance may be commercially damaging.

Notes

[1] "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)

[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] For example, see https://www.theguardian.com/technology/2018/feb/02/how-youtubes-algorithm-distorts-truth and YouTube's commitment to improve in 2019 https://blog.youtube/news-and-events/continuing-our-work-to-improve/.

[5] "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).

[6] "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); see also 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/), p. 54.

[7] The Facebook Papers, "What is Collateral Damage?", 12 August 2019, p. 34, cited in "Internal Facebook documents highlight its moderation and misinformation issues", TechCrunch, 25 October 2021 (URL: https://techcrunch.com/2021/10/25/internal-facebook-documents-highlight-its-moderation-and-misinformation-issues/)

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

[9] 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).

[10] "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.

[11] "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.

[12] 5x the amplification of a standard "like".

[13] "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/).

[14] "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.

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

[16] "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.

Irish Council for
Civil Liberties