

MARCH 2024

Safe by Default

Moving away from engagement-based rankings towards safe, rights-respecting, and human centric recommender systems

CONTENT

1
Executive Summary

4
Setting the context

8
A call for safe and quality-driven Recommender Systems by default

18
Conclusion

19
Endnotes

Executive Summary

Over the last decade, social media platforms have too often fallen short on their promise to connect and empower people and have instead become tools optimised to engage, enrage and addict them. The business model of the dominant platforms has created a profit incentive for platforms to prioritise user engagement over safety, with algorithmic recommender systems focused on keeping people clicking and scrolling as long as possible, which in turn allows the companies to sell more ad space, thereby generating revenue.

There is mounting evidence of the harms caused by ranking and recommending content being optimised for engagement. Ranking algorithms optimised for engagement select emotive and extreme content, and show it to people who they predict are most likely to engage with it (where “engage with” means they will scroll/stop scrolling to view or watch, click, reply, retweet, etc.). Meta's own internal research disclosed that a significant proportion (64%) of new joiners to extremist groups were caused by their own recommender systems. Even more alarmingly, in November 2023, Amnesty International found that TikTok's algorithms exposed multiple accounts of 13-year-old children to videos glorifying suicide within less than an hour of launching the account.

By determining how users find information and how they interact with all types of commercial and noncommercial content, recommender systems are a crucial design layer of Very Large Online Platforms (VLOPs)¹ regulated by the Digital Services Act (DSA).² Because of the specific risks they pose, recommender systems warrant urgent and special attention from regulators to ensure that platforms mitigate against “systemic risks”. Article 34 of the DSA defines “systemic risks” by reference to “actual or foreseeable negative effects” on the exercise of fundamental rights, dissemination of illegal content, civic discourse and electoral processes, public security and gender-based violence, as well as on the protection of public health and minors and physical and mental well-being.

As shown in our previous briefing, “Prototyping User Empowerment”, there are many ways for companies to mitigate against systemic risks, including by providing features that would encourage individuals to make conscious choices regarding content curation, promoting safer online behaviours and healthier habits.³ This transition towards authentic personalisation (i.e. an experience actively shaped by users) must start with VLOPs making their platforms safe-by-default. Unfortunately, this cannot be achieved with one quick switch. It will involve re-

ACKNOWLEDGEMENTS

This briefing was drafted by Katarzyna Szymielewicz (Panoptikon Foundation), with input from Tanya O'Carroll (independent expert), Marc Faddoul (AI Forensics), Dorota Głowacka (Panoptikon Foundation), and Oliver Marsh (AlgorithmWatch).

We would like to acknowledge valuable contributions and inspiration from the following experts:

Abigail Lawson, Integrity Institute

Claire Pershan, Mozilla Foundation

Jeff Allen, Integrity Institute

Johnny Ryan, Irish Council for Civil Liberties

Julian Jaurisch, Stiftung Neue Verantwortung (SNV)

Kasper Drazewski, BEUC, The European Consumer Organisation

Lisa Dittmer, Amnesty International

Margaux Vitre, École Normale Supérieure

Pat de Brún, Amnesty International

Rosie Morgan-Stuart, People vs Big Tech

Stanisław Burdziej, Nicolaus Copernicus University

Xavier Brandao, #jesuislà

designing many elements of the platform. This includes new features to actively promote more conscious user choice, opening up the social network infrastructure to third party content curation services, as well as measures aimed at protecting users from addictive and predatory design features.

In this briefing, we outline five categories of changes to the default settings of today's dominant social media platforms which will make their functioning safer, rights-respecting and human-centric:

1. Profiling off by default

In their default version VLOPs' recommender systems should not be based on behavioural profiling i.e. observing and collecting passive data about how users behave and interact on the platform in order to infer their interests. Instead, the default feed should only use as input signals data actively provided by the user for this very purpose (e.g. interests declared by the user when building their profile), as well as explicit user feedback on specific content (e.g. "show me more/show me less" signal sent by clicking a relevant button).

2. Optimising for values other than engagement

When designing their recommender systems VLOPs should depart from signals and metrics that correlate with user engagement (especially short term engagement) and prioritise signals/features that correlate with (subjective) relevance and (objective) credibility of the recommended content. This includes: prioritising the signals provided by explicit user feedback and preferences, bridging signals (e.g. the diversity of the users who engaged with a given piece of content and positive explicit feedback coming from users that are very different from one another), and signals that correlate with legitimacy, credibility and transparency of the source, especially when it comes to recommendations and search returns on sensitive topics.

3. Prompting conscious user choice, including opening up content curation to third party services

Platforms should create new features that facilitate conscious, authentic personalisation of the feed by their users and protect their wellbeing. This includes a range of measures such as sliders to set different optimization goals for recommendations (e.g. more long-form vs short-form content, local vs global relevance etc.), a 'hard stop' button to remove unwanted classifications of content from appearing altogether, a button to 'reset' an individual's feed, prompts to share declared interests and settings to allow users to explore how their feed changes based on their choices and interactions. A further promising avenue for user empowerment would be to oblige VLOPs to open up their infrastructure to allow independent, third-party content curation and moderation services.

4. Positive friction to disrupt compulsive behaviour and trigger reflection

Platforms should introduce positive friction aimed at slowing down posting and user interactions, giving users a chance to think before sharing. This includes 'think before you share' messages and limits on resharing as well as a series of practical recommendations aimed at countering platform 'stickiness' so that users are nudged towards disconnecting from social media rather than compulsively engaging, as well as being provoked to be more intentional about what they want to get out of a given social media session.

5. No addictive design features

Based on a growing body of research on the nature and impact of addictive design features on social media, we call on platforms to stop using certain design features altogether. These include measures like: notifications turned on by default, infinite scroll, video autoplay and misleading buttons which give users a false sense of control over content curation whilst not producing the results they advertise (such as “do not show content like this” buttons that do not prevent similar content from appearing again).



We appreciate that recommender systems are complex machines and any experimentation comes at a risk of causing new harms. Therefore measures recommended in this briefing should be tested and refined before implementation. This is the task for VLOPs guided by the European Centre for Algorithmic Transparency and the European Commission. What we hope is that, at the end of this process, (very large) social media platforms will have strong incentives to join a race to the top: competing with each other for default settings that prioritise safety and quality in user experience, and prototyping advanced features that allow for independent curation of recommended content.

Setting the context

Engagement-based ranking comes with social costs that can no longer be ignored

Tracking and profiling of users, and using this to power so-called personalised feeds which are optimised to keep users on platforms, has long raised concerns about harms for individuals and democratic societies.⁴ Research suggests that this technology drives social media addiction and poses mental health risks for users, in particular those with pre-existing vulnerabilities.

A key factor driving these concerns is **large social media platforms' choice to rank content primarily by the predicted probability of engagement**. Ranking algorithms optimised for engagement may prioritise emotive and extreme content, and show it to people who they estimate are most likely to engage with it (will watch, click, reply, retweet, etc.). According to Jonathan Stray, Jeff Allen and other contributors to "What We Know About Using Non-Engagement Signals in Content Ranking" study, engagement-based ranking disproportionately amplifies low-quality, misleading or sensational content that sparks a strong emotional reaction in the viewers rather than aiming to deliver them real value.⁵ A study published by Facebook itself has shown that content that comes closer to violating their terms of service, gets higher engagement and therefore greater amplification by their recommender systems.⁶

Amplification of divisive and polarising content may undermine social cohesion and push individuals towards political extremes. For example, Meta's own internal report revealed that a significant 64% of new joins to extremist groups were due to the platforms' recommendation algorithms.⁷ An investigation by Matthew Hindman, Nathaniel Lubin, and Trevor Davis in *The Atlantic* concluded that ranking algorithms reward 'superusers' who provoke the strongest engagement from others, usually reflected by the biggest number of interactions (irrespective of whether positive or negative), further skewing content on platforms towards divisive and controversial material.⁸

Driving user engagement also comes at the price of exploiting people's vulnerabilities and sensitive features (e.g. clicking and scrolling against their conscious intention not to look at certain content). In some cases this leads users into doomscrolling traps which negatively impact their wellbeing and may exacerbate pre-existing mental health issues (addictions, eating disorders, body complexes, anxiety or depressive disorders). Evidence on harmful consequences of the current functioning of recommender systems in this regard keeps emerging.⁹

In November 2023, an investigation by Amnesty International found that TikTok's algorithms exposed multiple 13-year-old child accounts to videos glorifying suicide in less than an hour of launching the account.¹⁰ A case study by Panoptykon Foundation (published in December 2023) showed that Facebook's recommender system ignored vulnerable user's explicit feedback, even when they requested to stop seeing certain content. Clicking on the "Hide post – See fewer posts like this" button 122 times on posts pertaining to illnesses, tragic accidents, and deaths

did not lead to lowering the frequency of such problematic content in the user's Facebook feed, with serious consequences on their mental health.¹¹

Detecting and addressing all harmful content, without inadvertently censoring legitimate content, is not an achievable goal. The use of machine learning requires working with certain margins of error, and unavoidable trade-offs between 'precision' and 'recall'. For example, the decision to weight an algorithm designed to detect harmful content towards greater 'precision' will result in more reliable detection of genuinely harmful content, but at the risk of not detecting less obvious cases (casting the net too narrowly). On the other hand, weighting an algorithm towards greater 'recall' will detect a wider range of harmful content, but at the risk of capturing false positives and inadvertently censoring legitimate content (casting the net too wide).

These limitations of algorithmic content moderation are yet another reason to demand that VLOPs monitor how harmful and 'borderline' content performs and redesign their systems in order to prevent such content from being amplified by recommender systems and in search results. The Center for Humane Technology made this call in their letter to Mark Zuckerberg,¹² following the release of the Facebook Files by whistleblower Frances Haugen in October 2021:

💡 *No matter how many fact-checkers you hire, how much you invest in AI, how you tweak metrics like Meaningful Social Interaction (MSI), or how hard your Oversight Board works, Facebook and democracy will be incompatible until the underlying operating model changes.¹³*

It is worth noting that Instagram has recently disabled their search engine for sensitive terms (related to suicide, self-harm and eating disorders).¹⁴ With this move Meta signalled that their algorithmic systems may be unsafe. But instead of fixing the search engine so that it returns safe results from trustworthy sources, they chose to turn it off. Apart from having a direct negative effect on users, algorithmic amplification of harmful and borderline content further incentivises its creation, as shown by Washington Post investigations in 2021.¹⁵

Last but not least, platforms' choices to optimise for content which is likely to engage users facilitates invasive harvesting and exploitation of users' personal data, in order to profile them for recommendations and sponsored content. Amnesty International argues that this is a serious and recurrent interference with people's right to privacy:

💡 *The sheer scale of the intrusion of Google and Facebook's business model into our private lives through ubiquitous and constant surveillance has massively shrunk the space necessary for us to define who we are. (...) The very nature of targeting, using data to infer detailed characteristics about people, means that Google and Facebook are defining our identity to the outside world, often in a host of rights-impacting contexts. This intrudes into our private lives and directly contradicts our right to informational self-determination, to define our own identities within a sphere of privacy.¹⁶*

Furthermore, recommender systems that feed on, and therefore process, sensitive data such as users' political views, sexual orientation, religion, ethnicity, or health information, may contravene the General Data Protection Regulation (GDPR).¹⁷ Processing 'special category' data is prohibited under Article 9(1) of the GDPR, apart from under certain circumstances which are unlikely to apply in the case of mass data harvesting.¹⁸

Recent developments in the EU

European Parliament's report on addictive design of online services and consumer protection in the EU single market

In December 2023, the European Parliament adopted, by a large majority, a report urging the Commission to enforce existing laws and urgently assess the need to prohibit the most harmful practices, such as infinite scroll, default autoplay, constant push and read receipt notifications, which are not yet blacklisted as misleading commercial practices.¹⁹ The report also calls for the Commission to assess potential addictive and mental health effects of engagement-based recommender systems, in particular hyper-personalised systems, that keep users on the platform as long as possible irrespective of whether this is good for the user.

MEPs underscored the need for providers of social media platforms to move away from features that focus on exploiting users' attention. They also stressed that policy actions in this area should "not place a burden on consumers, notably vulnerable users or their legal guardians, but address the harm caused by addictive design". MEPs called for more effective consumer protection through safer alternatives, even if these are not as profitable for social media platforms, and urged the Commission to "foster ethical design of online services by default" and "create a list of good practices of design features that are not addictive or manipulative."

21 civil society organisations and 9 distinguished academics against addictive design

In support of the European Parliament's Committee on Internal Market and Consumer Protection (IMCO)'s own initiative report on addictive design of online services and consumer protection in the EU single market,²⁰ a large group of civil society organisations, joined by distinguished experts in psychiatry, psychology, and computer science, sent an open letter to the European Parliament calling for legislative action against addictive design and other harmful features used by large online platforms:

💔💔 *We express our profound alarm at the social-media driven mental health crisis harming our young people and children. This is no glitch in the system. The platforms make more money the longer people are kept online and scrolling, and their products are therefore built around 'engagement at all costs' – leading to potentially devastating outcomes while social media corporations profit. Excessive and problematic social media use, such as compulsive or uncontrollable use, has been linked to sleep problems, attention problems, and feelings of exclusion among adolescents.²¹*

In the same letter, the People vs Big Tech Coalition called on the European Commission to:

- 💔💔 *[E]nsure strong enforcement of the Digital Services Act on the matter, with a focus on provisions on children and special consideration of their specific rights and vulnerabilities. This should include as a matter of priority:*
- *independently assessing the addictive and mental-health effects of hyper-personalised recommender systems;(…)*
 - *naming features in recommender systems that contribute to systemic risks; (…)*

- *examining whether an obligation not to use interaction-based recommendation systems 'by default' is required in order to protect consumers.*

MEPs calling for tech platforms' recommender systems to be switched off by default

On December 20, 2023, seventeen MEPs from various political groups (including S&D, the Left, Greens, EPP and Renew Europe) sent an open letter to Vice-President Vestager and Commissioner Breton urging the Commission to address concerns caused by engagement-based recommender systems:

💡 *Interaction-based recommender systems, in particular hyper-personalised systems, pose a severe threat to our citizens and our society at large as they prioritize emotive and extreme content, specifically targeting individuals likely to be provoked. (...) The insidious cycle exposes users to sensationalised and dangerous content, prolonging their platform engagement to maximise ad revenue.²²*

MEPs commended Ireland's new enforcer of the DSA and the Audiovisual Media Services Directive – Coimisiún na Meán – for moving toward effectively addressing the issues related to recommender systems. Acting on the basis of Article 6a(1) of the Audiovisual Media Services Directive (which empowers regulators to protect minors against potential harms) Coimisiún na Meán has issued a draft binding code for video platforms. It requires platforms such as YouTube and TikTok to ensure that “recommender algorithms based on profiling are turned off by default and that algorithms that engage explicitly or implicitly with special category data such as political views, sexuality, religion, ethnicity or health should have these aspects turned off by default.”²³

MEPs call upon the European Commission to follow Ireland's lead by recommending this measure as a mitigation measure to be taken by VLOPs in accordance with Article 35(1)(c) of the Digital Services Act.

A call for safe and quality-driven Recommender Systems by default

In this section we outline the five changes that will make social media platforms safer, rights-respecting and human-centric:

- No behavioural profiling by default
- Optimising for values other than engagement
- Prompting conscious user choice, including opening up content curation to third party services
- Positive friction to disrupt compulsive behaviour and trigger reflection
- No addictive design features

As shown in our previous briefing, social media platforms have many ways to help individuals make conscious choices regarding content curation, and develop safer online behaviours and healthier habits.²⁴ However, this transition towards authentic personalisation (experience actively shaped by users) and healthier habits must start with VLOPs changing their defaults and embedding different values into the design of their recommender systems.²⁵

As summarised by Stray et. al in their analysis of studies concerning recommender systems:

👉👉 *Even when controls are provided, many users do not know that they exist or what they do (...), find them challenging to use, or simply don't see the value in engaging with them. As a result, most users do not use recommender controls and a 'passive' user experience remains the default (...).*²⁶

A recent quantitative survey conducted in the EU has also shown that few consumers feel that they are in control over the content they see online, or even the choices that they make.²⁷ Their passive, resigned attitude is yet another outcome of what BEUC, the European Consumer Organisation, calls digital asymmetry and digital vulnerability:

👉👉 *Digital asymmetry is a term to describe how modern data-driven services put consumers at an unprecedented disadvantage. As they go online, they are faced with environments where traders control both the information that is presented and the entire choice architecture. (...) Even if consumers realise their online experience is personalised, they may never know the extent or mechanics of this personalisation, or the distortion it introduces into their view of the market or the world at large, and the choices they make as a result.*

*Digital vulnerability [is] a universal state of susceptibility to the exploitation of differences in power in the trader-consumer relationship resulting from internal and external factors beyond the control of the consumer. Such internal factors can include insufficient digital literacy, personal biases, limited cognitive capacity or plain information overload. External factors may include the digitally mediated relationship, the digital choice environments, the knowledge gap, limited control over data through user interfaces, the design of digital consumer environments, the lack of interoperability, the way default settings are configured, etc.*²⁸

In its work on digital fairness, BEUC argues that businesses benefiting from digital asymmetry should have a *positive* obligation to ensure ‘fairness by design’ – not only by protecting freedom of choice, but also by counteracting known biases (like clicking on the most prominent options) and by ensuring an environment where the individual does not need to make an effort to shield themselves from negative consequences (like finding and disabling tracking features, or dissecting a policy for potential risks).²⁹ Formulation of a positive principle-based obligation ‘to trade fairly’ has also been advocated for as a more flexible, cost-effective and more futureproof solution, rather than relying on identification of unfairness which is currently the European standard.³⁰

Against this backdrop, we argue that an essential and efficient way to prevent individual and societal harms resulting from the design and functioning of recommender systems is to change their default settings to the safest and most responsible version of the system.

Systemic risks cannot be mitigated without VLOPs changing their optimisation objective, which is reflected in the top line metrics used to measure recommender system performance.

Knowing that VLOPs make choices at each level of an algorithmic content curation system, we expect that they prioritise quality or relevance instead of engagement at all costs, which promotes divisiveness and excessive platform use.³¹

A major reason large online platforms continue to prioritise user engagement is the expectation of their shareholders to maintain growth and high profits, even at the expense of users’ interests.³² Now, with the DSA in place, we have reason to expect that shareholders’ interest will no longer prevail over fundamental rights of their users, public interest, and VLOPs’ social accountability.

Unfortunately, safer and human-centric functioning of recommender systems cannot be achieved with one switch. It will involve re-designing many elements of the platform. This includes new features to actively promote more conscious user choice, opening up the social network infrastructure to third party content curation services, as well as measures aimed at protecting users from addictive and predatory design features. Below we outline five categories of changes to the default settings of today’s dominant social media platforms, which will make their functioning safer, rights-respecting and human-centric:

❶ No behavioural profiling by default

The default versions of the recommender systems provided by platforms should not be based on behavioural profiling – i.e. observing and collecting passive data about how users behave and interact on the platform in order to infer their interests. **Instead, the default feed should only use data actively provided by the user for this very purpose as input signals (e.g. interests declared by the user when building their profile) as well as explicit feedback (e.g. “show me more/show me less” signal sent by clicking a relevant button).**

The demand to disable profiling-based recommender systems by default (as a way to mitigate systemic risks caused by these systems) has recently gained traction among expert civil society organisations. As argued by Amnesty International in their recent report “Driven into the Darkness” (which examines how TikTok’s “For You” feed encourages self-harm and suicidal ideation):

💡💡 *To respect privacy and to provide users with real choice and control, a profiling-free social media ecosystem should not just be an option but the norm. Content-shaping algorithms used by TikTok and other online platforms should therefore not be based on profiling (for example, based on watch time or engagement) by default and must require an opt-in instead of an opt-out, with the consent for opting in being freely given, specific, informed (including using child-friendly language) and unambiguous.³³*

The same argument was made by the Irish Council for Civil Liberties (ICCL) in their note “Ending amplification of hate & hysteria. Rapidly resolving the recommender system crisis”:

💡💡 *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.³⁴*

In February 2024, seventeen organisations supported this recommendation in their joint letter to Commissioner Thierry Breton.³⁵

In addition to mitigating societal risks related to civic discourse, electoral processes, physical and mental well-being, we see this measure as necessary and proportionate to protect personal data of individuals who continue to use large online platforms, in particular to prevent potentially unlawful processing of their sensitive data. Even though VLOPs argue that they do not intend to use sensitive characteristics when profiling their users and targeting content to them, users report that their vulnerabilities are being exploited.³⁶ As long as recommender systems use behavioural patterns to customise user experience, individual vulnerabilities (including sensitive characteristics) will be detected and (unintentionally) exploited by ML algorithms. Preventing profiling on protected attributes would require training an ML model to recognise (and protect) these characteristics, which in turn would require the platform to collect data on these characteristics from their users, which is certainly not what we would recommend as a mitigation measure.

Another argument against serving hyper-personalised content recommendations by default is that content curation algorithms should create feeds with content designed to serve a range of purposes, rather than being dominated by content predicted to keep a user engaged.

2 Optimising for values other than engagement

In addition to respecting privacy and vulnerabilities of their users, VLOPs need to reduce the prevalence of harmful clickbait and misinformation. As argued in the introduction to this part of the briefing, to do this VLOPs should change top line metrics used to measure recommender system performance and their definition of success (optimisation goal).

When designing their recommender systems VLOPs should depart from signals and metrics that correlate with user engagement (especially short term engagement) and prioritise signals/features that correlate with (subjective) relevance and (objective) credibility of the recommended content.

This could include:

- users' explicit feedback (user interaction with “show me less/show me more” features; results of surveys e.g. asking users to rate content's relevance or subjective value: “was your time watching this well spent?”);
- choices and preferences expressed by users (e.g. followed accounts, and names of accounts/publishers that users search, as a proxy of relevance for a given person);
- diversity of the users who engaged with a given piece of content³⁷ and other bridging signals, as defined by Aviv Ovadya and Luke Thorburn (e.g. positive explicit feedback coming from users that are very different from one another);³⁸
- established information retrieval signals (e.g. measuring traffic to the account's webpage from external domains, as per Google Search's PageRank);
- transparency of the actors behind the account and their history of original content creation.

In addition we recommend that VLOPs introduce prompts and tailored recommendations to prevent their users from:

- **falling into “doomscrolling traps”, e.g. excessive exposure to self-harm, diet-related content or idealised body images, which triggers “unhealthy” engagement and negatively impacts users' wellbeing,³⁹ and**
- **locking themselves in so-called “filter bubbles”, whereby personalised search results, recommendation systems and algorithmic curation isolate users in echo chambers where they only encounter information and opinions that conform to and reinforce their own beliefs.⁴⁰**

Measures discussed in this section should also have a mitigating effect on risks engagement-based recommender systems pose to civic discourse and electoral processes, as discussed by Sofia Calabrese (EPD) and Orsolya Reich (Liberties) in their recent paper. Among other mitigating measures, authors recommend that VLOPs:

- *develop algorithms that offer a balanced information diet, exposing users to a variety of viewpoints, particularly on controversial topics; (...)*
- *highlight content that receives positive responses across the political spectrum and present users with more content that resonates with a wide range of audiences from different groups; (...)*
- *provide links to accurate information.⁴¹*

We would like to see the following measures tested on social media platforms (and have results of these tests published to inform further discussion on mitigation measures):

- Introducing diversity as a metric of success for a recommender system and making sure that the stream of personalised recommendations includes a range of topics, instead of exploiting the one that is most engaging for the user in a given moment. For example, one recommendation out of ten should introduce a different topic selected from the range of what is predicted to be interesting for the user.
- When users search for information on public interest issues (such as conflicts, elections and politics, natural disasters, public health) VLOPs recommender systems should, in the first place, guide them to authoritative and trustworthy sources (e.g. public agendas such as WHO, national electoral commissions, publishers ranking high in NewsGuard ratings,⁴² certified according to Journalism Trust Initiative indicators⁴³ or Trust Project News Partners⁴⁴).

- When users make queries about sensitive topic areas, which include health (searches related to vaccines, medical treatment, diagnostics, epidemiology, coping with mental health issues), government/policy and financial advice, VLOPs recommender systems should, in the first place, guide them to authoritative sources (e.g. public services and helplines). If there aren't any authoritative results for a search query (e.g. because it relates to a conspiracy theory) a recommender system should return content on related and adjacent topics from authoritative sources.

3 Prompting conscious user choice, including opening up content curation to third party services

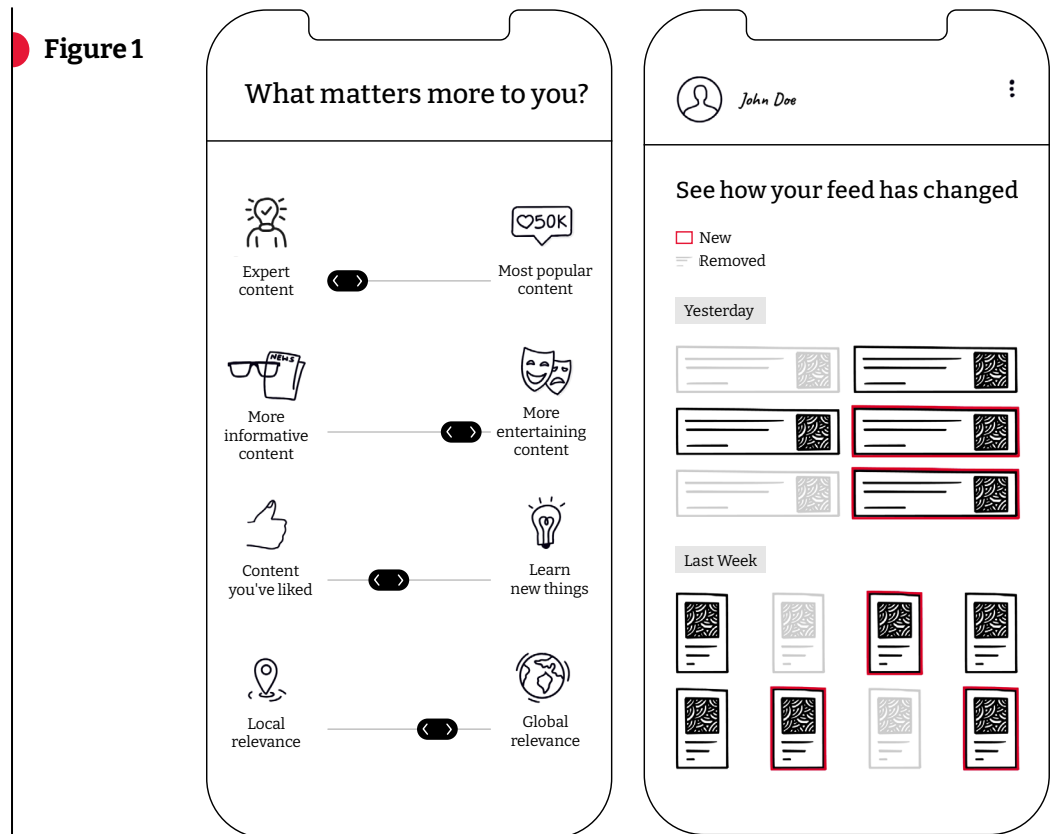
Acknowledging that some social media users value personalisation (and won't be satisfied with a feed organised chronologically or according to what is "trending" in a given area)⁴⁵ our recipe for safer, human-centric recommender systems is premised on a much more central role for conscious user choice and empowerment. While forced and engagement-oriented 'personalisation' of the feed leads to many documented harms, we advocate for features that lead to authentic personalisation of the feed when it is initiated and controlled by the user. **By default VLOPs should promote these features and educate their users about ways to customise their experience.**

As explained in our previous briefing, we call on VLOPs to create new features that facilitate conscious, authentic personalisation of the feed by their users and protect their wellbeing. Such features should also be seen as risk mitigation measures required under Article 35 of the DSA. These may include:

- sliders to set preferences for content curation (such as 'give me more informative content' or 'give me more entertaining content');
- a 'hard stop' button to suppress future recommendations of content related to a specific topic, hashtag or user (based on this signal, content curation algorithm should filter out content on topics flagged by the user as undesired);
- a button to reset individual feed (users who are concerned their feeds have become toxic should be able to reset the algorithm so that it discards all profiling information);
- feature allowing users to verify how their explicit feedback influenced selection of content that has been recommended to them;
- prompts to communicate their actual interests or pre-define their preferences about the type of content they wish (or NOT wish) to see (for example, users could be asked to enter specific interests if and when they opt-in for personalised recommendations);
- prompts to learn about factors that affect the content they see and post on a service (information about the signals and features used to rank organic and advertising/sponsored content as well as about reasons why user-generated content may be down-ranked).⁴⁶

Here we want to stress the importance of **not only making feedback and control features user-friendly but also making sure that they bring expected mitigating effects**. Providing deceptive or ineffective features which, from a user perspective, do not change their experience in a positive way, should be seen as a breach of Article 35 of the DSA (lack of effective mitigation measures against risks stemming from platform design), Article 27 (that users shall have transparent choices about recommender systems) and Article 25 (no deceptive design).⁴⁷ As previously mentioned, research has demonstrated the harmful consequences and user

See *Figure 1* below for examples of how these features might appear



frustration provoked by buttons that do not appear to work (in so far as they do not produce the advertised outcome). Such frustration may also explain why users tend to rely on default settings rather than attempt to exercise control over their feeds.

Acknowledging that choice can be burdensome and time-consuming, especially if executed on a daily basis, we advocate for solutions which take this burden off individuals' shoulders while still providing them meaningful control over algorithmically curated feeds. One very promising avenue for user empowerment would be to oblige VLOPs to open up their infrastructure to allow **independent, third-party content curation and moderation services**.

As summarised by Jean Cattani, Secretary General of the French Digital Council:


🗨️ *Opening social networks to third-party actors can enable them to introduce new value propositions to consumers, positively impacting information circulation (combating misinformation, protecting audiences, media pluralism, better access to information, and moving away from the attention economy). These third-party actors can offer alternative content recommendations, third-party applications, more advanced moderation forms, etc.⁴⁸*

Unbundling the social networks could address many of the harms connected to addictive design and predatory data surveillance by providing consumers with a marketplace of options for recommender systems and other content curation tools beyond the defaults offered by the Big Tech platforms today. This would also address the problematic nature of relying on VLOPs themselves as the arbiters of quality and credibility in ranking algorithms. Instead, there would be a marketplace of options, allowing for different regional and linguistic markets to be more adequately served, as well as new services to emerge that are exclusively serving the public interest (for example, users could choose to select a plug-in recommender system

operated by the public broadcaster in their country). Bluesky, a decentralised social network protocol launched by ex-Twitter CEO Jack Dorsey, is one example of what is technically possible.⁴⁹

We acknowledge that unbundling social media platforms will be a complex task, which merits a separate discussion paper. However, it is worth noting that there is already a significant body of academic literature to draw from, summarised in the box below.

Summary of academic literature

 *In 2020, a Working Group on Platform Scale at Stanford University proposed introducing middleware to tackle centralised platform power. They describe middleware as “software and services that would add an editorial layer between the dominant internet platforms and internet users”, following a definition used by Francis Fukuyama in “Making the internet safe for democracy”.⁵⁰ The intention of middleware is to dilute the concentrated power of the tech companies to control information flows on their platforms, and reduce the impact of algorithmic amplification. (...) Increased user agency and a decentralised middleware market could give more individual choice and power to users. This approach could generate greater competition among providers, and therefore dilute the impact of a small number of particularly powerful companies.⁵¹*

According to Daphne Keller, content-curation services give users more control over the material they see on internet platforms such as Facebook or Twitter. “Building on platforms’ stores of user-generated content, competing middleware services could offer feeds curated according to alternate ranking, labelling, or content-moderation rules.”⁵² At the same time Keller has acknowledged that unbundling social media platforms may pose long-term challenges involving privacy. To prevent this scenario, in “Getting Privacy Right” Keller proposed technical solutions, which should be discussed and implemented in early stages of such transformation.⁵³

Oliver Marsh draws attention to another risk related to independent content-curation services: enabling “truly personalised rabbit holes” for users who seek such experience.⁵⁴ While acknowledging this problem, we argue that it already exists in a monopolised social media environment and won’t be solved as long as feed personalisation is allowed. However, prompts and positive friction discussed in further sections of this briefing can mitigate this risk to a certain extent (the same measures should be required from third-party content curation services).

There have been some, relatively small-scale, attempts to develop middleware solutions, such as Gobo 2.0⁵⁵ and Prosocial Ranking Challenge.⁵⁶ However, some of these have run into issues with larger platforms, for instance Block Party which was closed down because of changes to Twitter data access.

4 Positive friction to disrupt compulsive behaviour and trigger reflection

The ease and speed of frictionless posting and sharing content has contributed to systemic risks discussed in the first part of this briefing. It has been established that the reshare button contributes significantly to the spread of misinformation and other harmful content. As explained by the Center for Humane Technology in their campaign #OneClickSafer:

👄 *One-click, frictionless sharing removes any barriers of action. When it's so easy to share, thoughtfulness drops and reactivity rises. You just click "share". On Facebook, this allows misinformation, hate speech, violence, and nudity to spread. And because its algorithms prioritize content based on engagement, the most harmful and engaging content goes viral.⁵⁷*

As explained in The Wall Street Journal's investigative podcast based on the Facebook Files:

👄 *It sounds almost too simple but literally every single hop of a reshare, it gets worse. So if a thing's been reshared 20 times in a row, it's going to be 10X or more likely to contain nudity, violence, hate speech, misinformation, than a thing that has just been not reshared at all.⁵⁸*

The same argument has been made by Accountable Tech in their report "Democracy by design. A Content-Agnostic Election Integrity Framework for Online Platforms":

👄 *Frictionless resharing is a staple of social platforms – and a key driver of toxicity. Internal Meta research showed⁵⁹ users are 4x more likely to encounter falsehoods in a reshare of a reshare than in the News Feed in general, and concluded⁶⁰ aggressively limiting these 'deep reshares' would be "an effective, content-agnostic approach to mitigate the harms."⁶¹*

Positive friction aims to slow down posting and user interactions, giving users a chance to think before sharing. Such intention could be incorporated into algorithmic ranking and when designing user interfaces. Based on their critical study of TikTok's "For You" (algorithmically curated) feed, Amnesty International calls for 'friction' measures as a mitigation strategy: VLOPs should "incorporate measures to limit the rapid and often disproportionate algorithmic amplification of borderline content".⁶²

Below we list examples of how positive friction can be introduced in recommender systems:

'What are you here for?'. A teaser to learn about a user's intention.⁶³

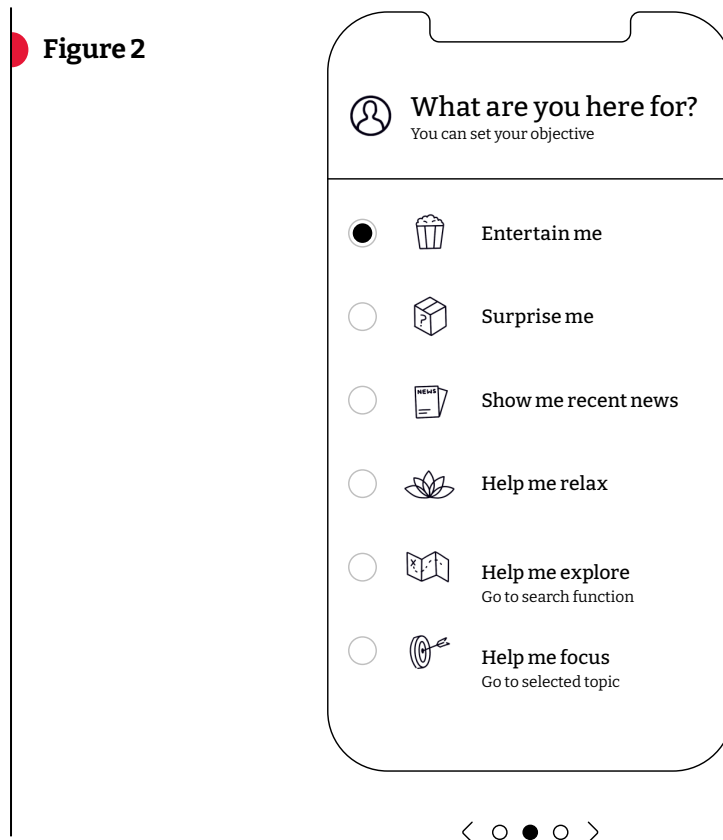
Showing an 'onboarding' screen every now and then would add friction to enter the service, therefore disrupting a habit of opening an app. It would make the user stop and reflect 'what is my intention for this session?'. In many cases, the answer to that question might be 'nothing really' or 'procrastination' (which is fine, as long as it is a choice and not a compulsive behaviour encouraged by addictive design features). This short moment of reflection may lead some users to leave the app if it would not actually be beneficial to them at that time.

Friction to share content

Based on the research by the Center for Humane Technology, Accountable Tech, Amnesty International and Frances Haugen testimony quoted above⁶⁴, we recommend the following measures be implemented by VLOPs:

- introducing 'think before you share' prompts;
- introducing a limit of reshares to slow down high virality content⁶⁵ (once this limit has been reached, users can still share given content but only by transforming it into a new post or message);⁶⁶
- nudging users to perform the task of rating the accuracy before sharing.⁶⁷

See Figure 2 below for an example of how this onboarding screen might appear




Other examples of design features that add positive friction and may encourage healthier online behaviour:

- warnings when users have spent more than 30 minutes on a specific service;
- prompts to set own, granular time limits for using the service;
- if notifications are turned on, prompts to select only necessary notifications and a delivery method that maximises users' digital wellbeing (e.g. notifications of new private messages or replies to posts delivered only after a given time period/ at a time specified by the user);
- prompts to set a day off social media;
- automatic locks after a preset time of use and during hours set by the user as 'offline';
- warnings when a user attempts to change safe default settings (including an encouragement to learn more about risks this change may incur); VLOPs should also run in-app awareness campaigns on potential risks resulting from problematic online behaviours;
- slowing down the feed towards the evening;⁶⁸
- slowing down the feed proportionately to time a given user spent on the platform, so that excessive use becomes less rewarding (regardless of the time of the day);
- introducing a 'circuit breaker': a prompt for the user to search for new content after a given number of consecutive interactions with recommended content ("Is this really what you want to keep seeing?") or, in a more radical version, a hard stop for the recommended feed in a given session ("You have seen all personalised recommendations for your query. If you want to see new recommendations, re-enter or refine your query").

5 No addictive design features

Within the framework of the European Unfair Commercial Practices Directive, deceptive design practices are defined as designs that “materially distorts or is likely to materially distort the economic behaviour with regard to the product of the average consumer whom it reaches or to whom it is addressed, or of the average member of the group when a commercial practice is directed to a particular group of consumers.”⁶⁹ The identification and prohibition of these practices have been pillars of consumer law for decades. Thus, recommender systems used by very large social media platforms to rank user-generated and sponsored content, are not exempt from it. However, social media platforms reliance on real-time micro-targeted tracking, complexifies the identification of deceptive design practices.⁷⁰ Hence, there is an urging need to adapt identification methods to social media platforms.

Indeed, in attention-based economy technology companies use design and system functionalities to take advantage of users’ and consumers’ vulnerabilities in order to capture their attention and increase the amount of time they spend on digital platforms. This claim is especially true for dominant social media platforms that chose engagement as their primary business objective. According to the European Parliament’s report on addictive design of online services and consumer protection in the EU single market:

 *Recommender systems, which are based both on personalisation and on interaction such as clicks and likes, potentially represent an important persuasive, addictive or behavioural design feature; (...) simultaneously recommender systems can contribute to the functionality of platforms to enhance social interaction, but are often also aimed at keeping users on the platform.⁷¹*

With these alarming concerns in mind, researchers and UI/UX designers are coming up with methods to identify deceptive design practices, which can potentially enhance addictive behaviour.⁷² The European Data Protection Board has already named a range of deceptive design features such as decontextualisation or ambiguous wording that should be easy for platforms to address immediately.⁷³

Data made available to researchers through the DSA should hopefully allow for further academic investigation into the nature and impact of addictive design features on social media users.⁷⁴ Based on the existing body of research,⁷⁵ we call on VLOPs to **stop using the following design features:**

- notifications turned on by default,
- infinite scroll,
- autoplay function,
- counts on social validation signals (such as like/dislike button),
- ‘fake’ buttons (giving users a false sense of control over content curation and, as a result, leading to their disappointment or real harms),⁷⁶
- complicated user paths to transparency or contestability features.⁷⁷

Conclusion

Engagement-based ranking comes with social costs that can no longer be ignored. The DSA has rightfully highlighted “systemic risks”⁷⁸ stemming from digital technologies to fundamental rights, civic discourse and electoral processes, and the protection of public health (physical and mental well-being). Expert civil society organisations and a host of researchers, referenced throughout this briefing, agree that a key factor driving such risks is large social media platforms’ choice to rank content primarily by the predicted probability of engagement.

Harms and negative effects related to the functioning of recommender systems have been researched for more than a decade, with plenty of evidence that today’s social media ecosystem is fundamentally broken. Very large online platforms tried to avoid their social responsibility by shifting the burden of proof to the victims of their exploitative business model. But thanks to whistleblowers and insiders who left these companies, we know that their owners are well aware of the addictive potential of their services and other harms experienced by their users. If amplification of borderline content in engagement-based recommender systems is not a bug, but a feature, the only reasonable response of the market regulator to VLOPs is: “change your default settings to prevent this negative effect!”

In this briefing we explained what it will take in practice to design safer and human-centric recommender systems. We stressed that this change needs to start with the optimisation objective, reflected in how VLOPs measure their recommender systems’ performance. Knowing that VLOPs make choices at each level of an algorithmic content curation system, we demand that they prioritise safety and quality of user experience (instead of engagement at all costs).

With the DSA in place, we expect that shareholders’ interest will no longer prevail over fundamental rights of social media users and VLOPs’ social accountability. As shown by a long list of references and publications quoted in this briefing, this expectation is backed by a growing coalition of civil society organisations and research institutions. With no time to waste, we call on the European Commission to use its powers under the DSA to define safe defaults for (very large) social media platforms and ensure that deceptive and harmful design features disappear from the market.

We appreciate that recommender systems are complex machines and any experimentation with their rules comes at a risk of causing new harms. Therefore we stress that measures recommended in this briefing should be tested before implementation. This is not something that we can do, as ‘adversarial auditors’ and independent researchers. This is a task for VLOPs, supervised and guided by the European Commission. But we reiterate our commitment to support this process in collaboration with the European Centre for Algorithmic Transparency.

1. This is a legal term under the European Union's Digital Services Act (DSA) which applies to platforms which have more than 45 million users per month in the EU. A full list can be found here: <https://digital-strategy.ec.europa.eu/en/policies/dsa-vlops>.
2. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act). Official text can be found here: <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R2065>.
3. Jesse McCrosky, Tanya O'Carroll, Caroline Sindors, Katarzyna Szymielewicz, "Prototyping User Empowerment", accessed February 19, 2024, https://panoptykon.org/sites/default/files/2023-11/peoplesbigtech_panoptykon_prototyping-empowerment_brief_20112023.pdf.
4. Personalisation has been a promise made by social media platforms to their users. We argue that as long as algorithmic content curation is optimised for user engagement (being an overarching commercial objective), the outcome (recommended content) is not 'personalised', only 'targeted'.
5. Tom Cunningham, Sana Pandey, Leif Sigerson, Jonathan Stray, Jeff Allen, Bonnie Barrilleaux, Ravi Iyer, Smitha Milli, Mohit Kothari, Behnam Rezaei, "What We Know About Using Non-Engagement Signals in Content Ranking", last modified February 5, 2019, <https://arxiv.org/ftp/arxiv/papers/2402/2402.06831.pdf>.
6. See Katarzyna Szymielewicz, Dorota Głowacka, "Fixing Recommender Systems. From identification of risk factors to meaningful transparency and mitigation", accessed February 2, 2024, p. 3-5. [panoptykon_ICCL_PvsBT_Fixing-recommender-systems_Aug 2023.pdf](https://panoptykon.org/sites/default/files/2023-12/panoptykon_algorithms-of-trauma-2-case-study-report_dec-2023.pdf).
7. Jeff Horwitz, "Facebook Executives Shut Down Efforts to Make the Site Less Divisive", *The Wall Street Journal*, last modified May 26, 2020, <https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499>.
8. Researchers found that the top 1% of accounts were responsible for 35% of all observed interactions; the top 3% were responsible for 52%. These hyper influential users were also the most abusive, skewing the publicly available inventory towards borderline content. See Matthew Hindman, Nathaniel Lubin, and Trevor Davis, "Facebook Has a Superuser-Supremacy Problem", *The Atlantic* online, last modified February 10, 2022, <https://www.theatlantic.com/technology/archive/2022/02/facebook-hate-speech-misinformation-superusers/621617/>.
9. See e.g. the collected works in Jonathan Haidt, Jean Twenge, Zach Rausch, "Social media and mental health: A collaborative review. Unpublished manuscript", New York University, accessed February 27, 2024, tinyurl.com/SocialMediaMentalHealthReview, as well as the "Surgeon General Issues New Advisory About Effects Social Media Use Has on Youth Mental Health", US Department of Health and Human Services, accessed February 27, 2024, <https://www.hhs.gov/about/news/2023/05/23/surgeon-general-issues-new-advisory-about-effects-social-media-use-has-youth-mental-health.html>.
10. "Global: TikTok's 'For You' feed risks pushing children and young people towards harmful mental health content", last modified November 7, 2020, <https://www.amnesty.org/en/latest/news/2023/11/tiktok-risks-pushing-children-towards-harmful-content/>.
11. Dorota Głowacka, Katarzyna Szymielewicz, Piotr Sapieżyński, "Algorithms of Trauma #2", accessed February 19, 2024, https://panoptykon.org/sites/default/files/2023-12/panoptykon_algorithms-of-trauma-2-case-study-report_dec-2023.pdf.
12. "To: Mark Zuckerberg", Center For Humane Technology online, accessed February 19, 2024, <https://www.humanetech.com/oneclicksafer-letter>.
13. "The Facebook Files" *The Wall Street Journal* online, accessed February 19, 2024 <https://www.wsj.com/articles/the-facebook-files-11631713039>.
14. "Hiding More Results in Instagram Search Related to Suicide, Self-Harm and Eating Disorders", *Meta* online, last modified January 9, 2024, <https://about.fb.com/news/2024/01/teen-protections-age-appropriate-experiences-on-our-apps/>.
15. Jeremy B. Merrill, Will Oremus, "Five points for anger, one for a 'like': How Facebook's formula fostered rage and misinformation", *Washington Post* online, last modified October 26, 2021, <https://www.washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/>. According to this investigation, European political parties stated that Facebook's ranking algorithms forced them to use "far more negative content than before", because engagement on positive and policy posts had fallen dramatically. See also Loveday Morris, "In Poland's politics, a 'social civil war' brewed as Facebook rewarded online anger", *The Washington Post* online, accessed February 21, 2024, <https://www.washingtonpost.com/world/2021/10/27/poland-facebook-algorithm/>.
16. "Amnesty International: Surveillance giants: how the business model of Facebook and Google threatens human rights", *Amnesty International*, last modified November 21, 2019, p. 22, <https://www.amnesty.org/en/documents/pol30/1404/2019/en/>. See also "Amnesty International: 'I Feel Exposed': Caught in TikTok's Surveillance Web", last modified November 7, 2023, <https://www.amnesty.org/en/documents/POL40/7349/2023/en/>.
17. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official text can be found here: <https://eur-lex.europa.eu/eli/reg/2016/679/oj>.
18. "Joint submission on the draft Online Safety Code", accessed February 21, 2024, p.5, https://www.iccl.ie/wp-content/uploads/2024/01/submission-60-civil-society-organisations-Coimisiun-na-Mean_

OSC-Consultation-Response.pdf.

19. “Report on addictive design of online services and consumer protection in the EU single market”, accessed February 19, 2024, https://www.europarl.europa.eu/doceo/document/A-9-2023-0340_EN.html.

20. “Addictive design of online services and consumer protection in the EU single market”, accessed February 19, 2024, [https://oeil.secure.europarl.europa.eu/oeil/popups/ficheprocedure.do?lang=en&reference=2023/2043\(INI\)](https://oeil.secure.europarl.europa.eu/oeil/popups/ficheprocedure.do?lang=en&reference=2023/2043(INI)).

21. “Open letter to the European Parliament: A critical opportunity to protect children and young people”, accessed February 19, 2024, <https://peoplevsbig.tech/open-letter-to-the-european-parliament-on-the-addictive-design-of-online-services>.

22. Johnny Ryan (@johnnyryan), “MEPs write to European Commission @ ThierryBreton & @Vestager asking for Irish @CNaM_ie rules against Big Tech's toxic algorithms to be applied across the EU.”, Twitter, December 20, 2023, <https://twitter.com/johnnyryan/status/1737415935772749882>.

23. “Consultation Document: Online”, Coimisiún na Meán, December 8, 2023, https://www.cnam.ie/wp-content/uploads/2023/12/Draft_Online_Safety_Code_Consultation_Document_Final.pdf.

24. McCrosky, O’Carroll, Sindere, Szymielewicz, “Prototyping User...”.

25. See recent studies discussing the failure of user control tools designed by social media platforms: Kelsey Smith, Georgia Bullen, Melissa Huerta, “Dark Patterns in User Controls: Exploring YouTube’s Recommendation Settings”, last modified November 30, 2021, <https://simplysecure.org/blog/dark-patterns-in-user-controls-exploring-youtubes-recommendation-settings/>.

See also Dietmar Jannach, Sidra Naveed, Michael Jugovac, “User Control in Recommender Systems: Overview and Interaction Challenges” In: Derek Bridge, Heiner Stuckenschmidt (eds) “E-Commerce and Web Technologies.

EC-Web 2016. Lecture Notes in Business Information Processing”, vol 278. Springer, Cham, February 15, 2017, https://doi.org/10.1007/978-3-319-53676-7_2.

26. Jonathan Stray and others, “Building Human Values into Recommender Systems: An Interdisciplinary Synthesis”, accessed February 19, 2024, <https://arxiv.org/ftp/arxiv/papers/2207/2207.10192.pdf>.

27. In a 2023 EU survey, less than one in two (43%) respondents reported that they feel in full control of the online content they are shown and the decisions they make online. See “Consumer survey results on the fairness of the online environment”, The European Consumer Organisation, accessed February 19, 2024, https://www.beuc.eu/sites/default/files/publications/BEUC-X-2023-113_Fairness_of_the_digital_environment_survey_results.pdf.

28. “Protecting fairness and consumer choice in a digital economy”, The European Consumer Organisation, February 10, 2022, https://www.beuc.eu/sites/default/files/publications/beuc-x-2022-015_protecting_fairness_and_consumer_choice_in_a_digital_economy.pdf.

29. “Towards European Digital Fairness. BEUC framing response paper for the REFIT consultation”, The European Consumer Organisation, February 20, 2023, https://www.beuc.eu/sites/default/files/publications/BEUC-X-2023-020_Consultation_paper_REFIT_consumer_law_digital_fairness.pdf.

30. See Paolo Siciliani, Christine Riefa, Harriet Gamper, “Consumer Theories of Harm: An Economic Approach to Consumer Law Enforcement and Policy Making”, last modified June, 2019, https://www.researchgate.net/publication/330798207_Consumer_Theories_of_Harm_An_Economic_Approach_to_Consumer_Law_Enforcement_and_Policy_Making/link/5c5463a8299bf12be3f3e32a/download.

31. See Integrity Institute, “On Risk Assessment and Mitigation for Algorithmic Systems”, February 2024,

<https://drive.google.com/file/d/1ZMt7igUcKUq00yakCnbxBCcaA7vajAix/view>.

32. See “Ranking by Engagement”, accessed February 19, 2024, <https://integrityinstitute.org/blog/ranking-by-engagement>.

33. “Driven into Darkness: How TikTok’s ‘For You’ Feed Encourages Self-Harm and Suicidal Ideation”, Amnesty International, last modified November 7, 2023, <https://www.amnesty.org/en/documents/pol40/7350/2023/en/>.

34. “The European Commission must follow Ireland’s lead, and switch off Big Tech’s toxic algorithms”, Irish Council for Civil Liberties, last modified December 18, 2023, <https://www.iccl.ie/2023/the-european-commission-must-follow-irelands-lead-and-switch-off-big-techs-toxic-algorithms/>.

35. People vs BigTech, “Letter to European Commissioner Breton: Tackling harmful recommender systems”, last modified February 5, 2024, <https://peoplevsbig.tech/letter-to-european-commissioner-breton-tackling-harmful-algorithms>.

36. See “Hypothesis 4” in Szymielewicz, Głowacka “Fixing Recommender Systems”.

37. Metric discussed in Cunningham and others, “What We Know...” p. 7.

38. Aviv Ovadya, Luke Thorburn, “Bridging Systems: Open problems for countering destructive divisiveness across ranking, recommenders, and governance”, last modified October 26, 2023, <https://knightcolumbia.org/content/bridging-systems>.

39. See Szymielewicz, Głowacka, “Fixing...” p. 9, for further discussion, accessed February 19, 2024, https://panoptykon.org/sites/default/files/2023-08/Panoptykon_ICCL_PvsBT_Fixing-recommender-systems_Aug%202023.pdf.

40. “Filter bubble”, Wikipedia, accessed February 19, 2024, https://en.wikipedia.org/wiki/Filter_bubble.

41. See Sofia Calabrese (EPD), Orsolya Reich (Liberties), “Identifying,

- analysing, assessing and mitigating potential negative effects on civic discourse and electoral processes: a minimum menu of risks very large online platforms should take heed to", accessed February 19, 2024, https://dq4n3btxmr8c9.cloudfront.net/files/mpdgy5/DSA_Risk_Analysis_LibertiesxEPDfin.pdf.
42. See "Website Rating Process and Criteria", accessed February 19, 2024, <https://www.newsguardtech.com/ratings/rating-process-criteria/>
43. See "JTI. Key indicators", accessed February 27, 2024, <https://rsf.org/en/journalism-trust-initiative#key-indicators-2923>.
44. See "The 8 Trust Indicators", accessed February 19, 2024, <https://thetrustproject.org/trust-indicators/>.
45. Solution offered by YouTube and TikTok as one of the alternatives to personalised feed.
46. McCrosky, O'Carroll, Sinderson, Szymielewicz, "Prototyping user...", p. 5.
47. See recommendations in Dorota Głowacka, Katarzyna Szymielewicz, Piotr Sapieżyński, "Algorithms of Trauma #2. Stuck in a 'doomscrolling trap' on Facebook? The platform will not let you escape", accessed February 20, 2024, https://panoptykon.org/sites/default/files/2023-12/panoptykon_algorithms-of-trauma-2_case-study-report_dec-2023.pdf.
48. French Digital Council, "Fostering the wealth of networks", February, 2024, p. 8, https://cnnumerique.fr/files/uploads/2024/Fostering_the_wealth_of_networks_French_digital_Council.pdf.
49. Amanda Silberling, Alyssa Stringer, Cody Corral, "TechCrunch, What is Bluesky? Everything to know about the app trying to replace Twitter", accessed February 19, 2024, <https://techcrunch.com/2023/11/17/what-is-bluesky-everything-to-know-about-the-app-trying-to-replace-twitter/>.
50. Francis Fukuyama, "Making the Internet Safe for Democracy", *Journal of Democracy*, vol. 32, April 2020, <https://www.journalofdemocracy.org/articles/making-the-internet-safe-for-democracy/>.
51. Excerpts from ISD paper of Sara Bundtzen, "Suggested for You: Understanding How Algorithmic Ranking Practices Affect Online Discourses and Assessing Proposed Alternatives", accessed February 19, 2024, <https://www.isdglobal.org/wp-content/uploads/2022/12/Understanding-How-Algorithmic-Ranking-Practices-Affect-Online-Discourses-and-Assessing-Proposed-Alternatives.pdf>.
52. Daphne Keller, "The Future of Platform Power: Making Middleware Work", *Journal of Democracy*, vol. 32, July 2021, <https://www.journalofdemocracy.org/articles/the-future-of-platform-power-making-middleware-work/>.
53. Daphne Keller, "Privacy, Middleware, and Interoperability: Can Technical Solutions, Including Blockchain, Help Us Avoid Hard Tradeoffs?", August 23, 2021, <https://cyberlaw.stanford.edu/blog/2021/08/privacy-middleware-and-interoperability-can-technical-solutions-including-blockchain-0>.
54. Oliver Marsh, "Social Media Futures: Interventions Against Online Unpleasantness", last modified April 19, 2022, <https://www.institute.global/insights/tech-and-digitalisation/social-media-futures-interventions-against-online-unpleasantness>.
55. Lane Spencer, "Gobo 2.0: All Your Social Media in One Place", last modified November 9, 2022, <https://publicinfrastructure.org/2022/11/09/gobo-2-0-all-your-social-media-in-one-place/>.
56. "The Prosocial Ranking Challenge – \$60,000 in prizes for better social media algorithms", last modified January 18, 2024, <https://humancompatible.ai/news/2024/01/18/the-prosocial-ranking-challenge-60000-in-prizes-for-better-social-media-algorithms/>.
57. "Make Facebook #OneClickSafer. Frictionless sharing is dangerous. Changing the reshare button is a proven solution", Center for Human Technology, accessed February 19, 2024, <https://www.humanetech.com/oneclicksafer>.
58. "The Facebook Files, Part 4: The Outrage Algorithm", *The Journal*, last modified September 18, 2021 <https://www.wsj.com/podcasts/the-journal/the-facebook-files-part-4-the-outrage-algorithm/e619fbb7-43b0-485b-877f-18a98ffa773f>.
59. Alex Kantrowitz, "The Case To Reform The Share Button, According To Facebook's Own Research", last modified November 5, 2021, <https://www.bigtechnology.com/p/the-case-to-reform-the-share-button>.
60. "Facebook misled investors and the public about its role perpetuating misinformation and violent extremism relating to the 2020 election and January 6th insurrection", accessed February 19, 2024, https://facebookpapers.com/wp-content/uploads/2021/11/Insurrection_Redacted.pdf.
61. "Democracy By Design. A Content-Agnostic Election Integrity Framework for Online Platforms", September 2023, <https://accountabletech.org/wp-content/uploads/Democracy-By-Design.pdf>.
62. "Driven into Darkness...", p. 66.
63. We have recommended this feature in our previous briefing: McCrosky, O'Carroll, Sinderson, Szymielewicz, "Prototyping user...", p. 10.
64. Frances Haugen advocated for limiting the number of times a piece of content can be shared to help reduce the disproportionate influence of superuser activity.
65. For example, WhatsApp has experimented with a limit of five forwards. After five forwards users can only forward the message to one person at a time. See Casey Newton, "WhatsApp puts new limits on the forwarding of viral messages", last modified April 7, 2020, <https://www.theverge.com/2020/4/7/21211371/whatsapp-message-forwarding-limits-misinformation-coronavirus-india>.
66. This concept has been promoted by the Center for Humane Technology

- in their campaign. More informations can be found here: "Make Facebook #OneClickSafer. Frictionless sharing is dangerous. Changing the reshare button is a proven solution", accessed February 20, 2024, <https://www.humanetech.com/oneclicksafer>.
67. Researchers from MIT Sloan observed that nudges to get users thinking about the accuracy of a piece of content made them more discerning when it came to sharing true or false news. Notably, users who performed the task of rating the accuracy first were less likely to share inaccurate news, and more likely to share accurate. See Bundtzen, "Suggested for You...", p. 22.
68. According to Frances Haugen, slowing down algorithmic ranking towards the evening could help incentivise superusers to switch off earlier, so algorithms receive fewer toxic signals from such users. See Bundtzen, "Suggested for You...", p. 22.
69. Directive 2005/29/EC of the European Parliament and of the Council of 11 May 2005 concerning unfair business-to-consumer commercial practices in the internal market and amending Council Directive 84/450/EEC, Directives 97/7/EC, 98/27/EC and 2002/65/EC of the European Parliament and of the Council and Regulation (EC) No 2006/2004 of the European Parliament and of the Council (Unfair Commercial Practices Directive) (Text with EEA relevance).
70. Lauren E. Willis, "Deception by design", *Harvard Journal of Law & Technology*, 2020, vol. 34, p. 115.
71. "Report on addictive design...", recital M.
72. Zewei Shi, Ruoxi Sun, Jieshan Chen, Jiamou Sun, Minhui Xue, "The Invisible Game on the Internet: A Case Study of Decoding Deceptive Patterns", last modified February 5, 2024, <https://arxiv.org/abs/2402.03569>.
73. "Guidelines 03/2022 of the European Data Protection Board on deceptive design patterns in social media platform interfaces: how to recognise and avoid them", February 14, 2023, https://edpb.europa.eu/system/files/2023-02/edpb_03-2022_guidelines_on_deceptive_design_patterns_in_social_media_platform_interfaces_v2_en_0.pdf.
74. Ref. to Article 40 of the DSA.
75. See e.g. the following publications: Anshul Vikram Pandey, Katharina Rall, Margaret L. Satterthwaite, Oded Nov, Enrico Bertini "How Deceptive are Deceptive Visualizations? An Empirical Analysis of Common Distortion Techniques. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems", last modified April 18, 2015, <https://dl.acm.org/doi/abs/10.1145/2702123.2702608>; Monge Roffarello, A., Lukoff, K., & De Russis, "Defining and Identifying Attention Capture Deceptive Designs in Digital Interfaces. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems", April, 2023, p. 1-19; National Academies of Sciences, Engineering, and Medicine; Health and Medicine Division; Board on Population Health and Public Health Practice; Committee on the Impact of Social Media on Adolescent Health; Sandro Galea, Gillian J. Buckley, Alexis Wojtowicz, "Social Media and Adolescent Health", National Academies, 2023, <https://nap.nationalacademies.org/catalog/27396/social-media-and-adolescent-health>.
76. Recent studies discussing consequences of misleading design in user feedback settings: Dorota Głowacka, Aleksandra Iwańska, "Algorithms of trauma: new case study shows that Facebook doesn't give users real control over disturbing surveillance ads", September 28, 2021, <https://en.panoptikon.org/algorithms-of-trauma>; Głowacka, Szymielewicz, Sapieżyński, "Algorithms of Trauma #2. Stuck in..."; Ashlee Milton, Leah Ajmani, Michael Ann DeVito, Stevie Chancellor, "I See Me Here: Mental Health Content, Community, and Algorithmic Curation on TikTok, last modified April 19, 2023; "Driven into Darkness...", <https://www.amnesty.org/en/documents/POL40/7350/2023/en/>; Becca Ricks, Jesse McCrosky, "Does this button work? Investigating YouTube's ineffective user controls", last modified September 20, 2022, <https://foundation.mozilla.org/en/research/library/user-controls/report/>; Alexander Liu, Siqi Wu, Paul Resnick, "How to Train Your YouTube Recommender to Avoid Unwanted Videos", last modified August 2, 2023, <https://doi.org/10.48550/arXiv.2307.14551>.
77. Smith, Bullen, Huerta, "Dark Patterns...".
78. Concept introduced by Article 34 of the DSA.