FUNDACJA PANOPTYKON

Irish Council for Civil Liberties

People vs BIGTECH

BRIEFING NOTE

Fixing Recommender Systems

From identification of risk factors to meaningful transparency and mitigation

CONTENT

Executive Summary

From August 25th 2023 Europe's new Digital Services Act (DSA) rules kick in for the world's largest digital platforms, shaping the design and functioning of their key services. For the nineteen platforms that have been designated "Very Large Online Platforms" (VLOPs) and "Very Large Online Search Engines" (VLOSEs), there will be many new requirements, from the obligation to undergo independent audits and share relevant data in their transparency reports, to the responsibility to assess and mitigate against "systemic risks" in the design and implementation of their products and services. 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.

One of the major areas where platform design decisions contribute to "systemic risks" is through their recommender systems – algorithmic systems used to rank, filter and target individual pieces of content to users. By determining how users find information and how they interact with all types of commercial and noncommercial content, recommender systems became a crucial design-layer of VLOPs regulated by the DSA. Shadowing their rise, is a growing body of research and evidence indicating that certain design features in popular recommender systems contribute to the amplification and virality of harmful content, such as hate speech, misinformation and disinformation, addictive personalisation and discriminatory targeting in ways that harm fundamental rights, particularly the rights of minors. As such, social media recommender systems warrant urgent and special attention from the Regulator.

VLOPs and VLOSEs are due to submit their first risk assessments (RAs) to the European Commission in late August 2023. Without official guidelines from the Commission on the exact scope, structure and format of the RAs, it is up to each large platform to interpret what "systemic risks" mean in the context of their services – and to choose their own metrics and methodologies for assessing specific risks.

ACKNOWLEDGEMENTS

This brief was drafted by Katarzyna Szymielewicz (Senior Advisor at the Irish Council for Civil Liberties) and Dorota Głowacka (Panoptykon Foundation), with notable contributions from Alexander Hohlfeld (independent researcher), Bhargav Srinivasa Desikan (Knowledge Lab, University of Chicago), Marc Faddoul (AI Forensics) and Tanya O'Carroll (independent expert).

In addition we are grateful to the following civil society experts for their contributions:

Anna-Katharina Meßmer, Stiftung Neue Verantwortung (SNV)

Asha Allen, Centre for Democracy and Technology, Europe Office

Belen Luna, HateAid

Josephine Ballon, HateAid

Claire Pershan, Mozilla Foundation

David Nolan, Amnesty International

Fernando Hortal Foronda, European Partnership for Democracy

Jesse McCrosky, Mozilla Foundation/Thoughtworks

John Albert, AlgorithmWatch

Lisa Dittmer, Amnesty International

Martin Degeling, Stiftung Neue Verantwortung (SNV)

Pat de Brún, Amnesty International

Ramak Molavi Vasse'i, Mozilla Foundation

Richard Woods, Global Disinformation Index

In order to assist the Commission in reviewing the RAs, we have compiled a list of hypotheses that indicate which design features used in recommender systems may be contributing to what the DSA calls "systemic risks". Our hypotheses are accompanied by a list of detailed questions to VLOPs and VLOSEs, which can serve as a "technical checklist" for risk assessments as well as for auditing recommender systems.

Based on independent research and available evidence we identified six mechanisms by which recommender systems may be contributing to "systemic risks":

1. amplification of "borderline" content (content that the platform has classified as being at higher risk of violating their terms of service) because such content drives "user engagement";
2. rewarding users who provoke the strongest engagement from others (whether positive or negative) with greater reach, further skewing the publicly available inventory towards divisive and controversial content;
3. making editorial choices that boost, protect or suppress some users over others, which can lead to censorship of certain voices;
4. exploiting people's data to personalise content in a way that harms their health and wellbeing, especially for minors and vulnerable adults;
5. building in features that are designed to be addictive at the expense of people's health and wellbeing, especially minors;
6. using people's data to personalise content in ways that lead to discrimination.

For each hypothesis, we provide highlights from available research, which support our understanding of how design features used in recommender systems contribute to harms experienced by their users. However, it is important to note that researchers have been constrained in their attempts to verify causal relationships between specific features of recommender systems and observed harms by what data was made available to them either by online platforms or platforms' users. Because of these limitations external audits have spurred debates about the extent to which observed harms are caused by recommender system design decisions or by natural patterns in human behaviour.

It is our hope that risk assessments carried out by VLOPs and VLOSEs, followed by independent audits and investigations led by DG CONNECT, will end these speculations by providing data for scientific research and revealing specific features of social media recommender systems that directly or indirectly contribute to "systemic risks" as defined by Article 34 of the DSA.

In the second part of this brief (page 14) we provide a list of technical information that platforms should disclose to the Regulator, independent researchers and auditors to ensure that results of the risk assessments can be verified. This includes providing a high-level architectural description of the algorithmic stack as well as specifications of different algorithmic modules used in the recommender systems (type of algorithm and its hyperparameters; input features; loss function of the model; performance documentation; training data; labelling process etc).

Revealing key choices made by VLOPs and VLOSEs when designing their recommender systems would provide a "technical bedrock" for better design choices and policy decisions aimed at safeguarding the rights of European citizens online.

You can find a full glossary of technical terms used in this briefing on page 16.

SECTION 1

How design features used in recommender systems contribute to "systemic risks", as defined by the DSA

HYPOTHESIS 1

Amplification of illegal and borderline harmful content because it drives "engagement"

Description

"One of the biggest issues social networks face is that, when left unchecked, people will engage disproportionately with more sensationalist and provocative content. (....) At scale it can undermine the quality of public discourse and lead to polarization. (...) Our research suggests that no matter where we draw the lines for what is allowed, as a piece of content gets close to that line, people will engage with it more on average. (...) This is a basic incentive problem that we can address by penalizing borderline content so it gets less distribution and engagement." (Zuckerberg, 2021)

This quote shows that social media executives are aware of what researchers call the "natural engagement pattern" and its consequences for public debate. Nevertheless, they choose engagement as an overarching objective for ranking content and constructing algorithmic feeds because this has proven to be the most profitable objective to optimise for. According to Tom Cunningham, engagement is negatively related to quality (Cunningham, 2023). Content with the highest predicted engagement scores low in terms of quality and trustworthiness (attributes of quality used by the platforms include: withholding information, sensationalised language and engagement bait) (META, n.d.).

When training content moderation algorithms (e.g hate-speech classifiers), platforms prioritise precision over accuracy, which reduces the likelihood of false positives (content wrongfully flagged as violating a platform's policies). This design choice is motivated by commercial interests: since violent, hateful and sensational content proves engaging, platforms delay removing it until violation of their policy or law is evident. Insufficient accuracy means that harmful content remains available on the platform even if flagged by users and, because of the engagement-oriented design choices of recommender algorithms (combined with the "natural engagement pattern" described above), such content is also likely to be amplified.

Amplification of hateful and violent content also affects users' behaviour. People who belong to marginalised groups and those who are committed to socially relevant issues, such as journalists, politicians or activists, are more exposed to attacks and, as a result, discouraged from participation.[1]

1. In 2021, an EU-representative survey of 2.000 people between the ages of 18 and 80 commissioned by HateAid and The Landecker Digital Justice Movement found that alarming rates of internet users, especially women, experience hate and violence online and thus change their behaviour and withdraw from social media (HateAid, 2021).

> 71% of videos that were reported as harmful were recommended to viewers on the platform.

Supporting Evidence

A Mozilla report found that 71% of videos that were reported as harmful (referred to as 'Regret reports' on YouTube) were recommended to viewers on the platform. In total, recommended videos were 40% more likely to be reported as harmful than videos users found via specific searches. These videos often contained harmful content such as violence, misinformation, hate speech and scams (Mozilla, 2021). A subsequent research showed inefficiency of user control tools that were available to YouTube users (buttons like "Dislike" and "Don't Recommend Channel") – it turned out they did not effectively prevent "unwanted" recommendations (Ricks & McCrosky, 2022).

In 2022, Amnesty International investigated Meta's role in the human rights violations against Rohingya. The report found that Meta's algorithm proactively amplified and promoted content that incited violence, hatred and discrimination against the Rohingya, which they suggest substantially increased the risk of an outbreak of mass violence. Core features highlighted in the report are: the platforms's newsfeed, ranking, and recommendation algorithms. According to leaked internal Meta documents, a team within Meta received an "escalation" of an anti-Rohingya video by the extremist monk, U Wirathu, on an unknown date in 2020. The video was reported for violating community standards. Meta found that its algorithms had been actively promoting the video by Wirathu, who was well-known for hate speech. The investigation revealed that over 70% of the video's views had come from "chaining", which meant that the platform was actively recommending divisive and inciting content. According to Meta, "chaining" is an example of "non follower-based distribution" and refers to video content "which auto-plays after a video is complete and suggests what's "Up Next" to viewers (Amnesty International, 2022).

In an internal presentation from 2016 reviewed by the Wall Street Journal, a company researcher, Monica Lee, found that Facebook was not only hosting a large number of extremist groups but also promoting them to its users: "64% of all extremist group joins are due to our recommendation tools," the presentation said, predominantly thanks to the models behind the "Groups You Should Join" and "Discover" features (Seetharaman & Horwitz, 2020).

In the context of German elections in 2021, a civil society investigation documented undermoderation of illegal content and disinformation, as well as amplification of divisive content, e.g., via automated recommendations for political pages, groups and profiles spreading hate, violence and disinformation or by placing paid ads for said content. This mechanism particularly benefited right wing extremist parties (Reset & Hate Aid, 2022).

Questions for Online Platforms

1. What behavioural signals (e.g comment or emotional reaction, minutes spent on video) are taken into account by the recommender system (as input features) and what "weights" are attributed to them?
2. How does explicit user feedback (control tools such as the "not interested" button on TikTok and more sophisticated filtering tools that allow users to define hashtags and terms related to content they do not want to engage with) influence ranking?
3. Does opting for a reverse-chronological timeline (instead of an algorithmic feed) alter user retention rate and how long users stay on the platform?

4. What proportion of user reach and engagement results from algorithmic amplification (ref. to data on organic vs algorithmic consumption)?
5. Do the recommendation system algorithms promote low quality but statistically engaging content (e.g. content withholding information, linking to misinformation, using sensationalised language and engagement bait) in addition to recommending similar content to what the user reacted to?
6. Does the platform classify content that promotes outrage (e.g "angry" reactions), and promote this content to increase engagement?
7. As borderline content varies by location and time, how often is the model retrained and updated? How is accuracy ensured in an environment, where language and imagery change constantly? When does a post cross the threshold to be flagged, and when is this threshold changed for specific events?
8. What was the proportion of content flagged as "borderline" in the content with highest visibility on the platform in a given period of time?
9. What was the proportion of content flagged as "borderline" by the platforms' detection system in recommended content over a given period of time?

HYPOTHESIS 2

Rewarding users who provoke the strongest engagement from others (whether positive or negative), further skewing the publicly available inventory towards divisive and controversial content

Description

Researchers observe that ranking algorithms reward "meaningful interactions" (such as comments) and engaging content. This mechanism promotes superusers who have disproportionate influence over how ranking algorithms weigh what could be interesting to other users. When a platform rewards the "wrong" users for long enough, those users become very powerful, trapping platforms and their algorithms. Research shows that superusers tend to be most abusive, skewing the publicly available inventory towards divisive and controversial content.

Supporting Evidence

In 2021/22 Matthew Hindman, Nathaniel Lubin, and Trevor Davis (The Atlantic) investigated the phenomenon of superuser-supremacy on Facebook, a class of users that produce more likes, shares, reactions, comments, and posts than 99% of users in the United States. Researchers analysed 52 million users, looking at 500 US-run pages with the highest average engagement as well as the highest-interaction posts from more than 41,000 of the highest-membership US public groups. 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. Among a randomly selected sample of 30,000 users, focusing on the 219 accounts with at least 25 public comments, 68%spread misinformation, reposted in a spam-like way, published comments that were racist or sexist or anti-Semitic or anti-gay, or incited violence (Hindman et al., 2022).

In six out of seven analysed countries Tweets posted by Twitter accounts from

> (Meta's) new metric "systematically" rewarded users or groups that posted divisive, shocking, misleading and low-quality content.

the political right received more algorithmic amplification than the political left when studied as a group. Right-leaning news outlets in the US, as defined by the independent organisations, saw greater algorithmic amplification on Twitter compared to left-leaning news outlets (Huszár et al., 2021).

NYU's research suggests that one plausible reason for the greater amplification of right-leaning news outlets, was that conservative politicians are more likely than their peers to be "ratioed" (i.e. when a tweet receives more "'quotes"' and "'replies"' than simple retweets). High ratio often indicates that the tweet may be unpopular, or soliciting a negative reaction (eg. causes more outrage and division online). Algorithms may however interpret the unpopularity of such tweets as increased engagement, and therefore amplify them (Brown et al., 2021).

In 2018 Facebook introduced a new metric for its News Feed algorithm, using what it referred to as "meaningful social interactions" for ranking people's interactions on the platform, assigning different point values to things such as "likes", comments, posts etc. Facebook started to weigh emojis (including the angry emoji) five times higher than the like button in the mathematical calculation of the recommendation algorithms. According to documents revealed by Frances Haugen, the new metric "systematically" rewarded users or groups that posted divisive, shocking, misleading and low-quality content, which seemed to have a significant impact on the increase of spreading misinformation and violent content across the platform. As a result content producers were disincentivised from posting more nuanced and fact-based information. 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. The new metric was eventually rowed back - with the result being that users received less problematic content (Merrill & Oremus, 2021).

Questions for Online Platforms

1. With reference to input features (mentioned in the previous section), which interactions are most important (influential) for rating content? Does the platform use a classification like Meta's "meaningful interactions" and, if yes, how is it defined?
2. When collecting behavioural observations, does the platform differentiate between negative and positive engagement (e.g. a user forwarding content to their own followers, as a sign of approval, versus adding negative comments, as a sign of disapproval)? How does this classification influence rating and ranking content?
3. How many users are responsible for, respectively, 10%, 20% and 30% of all observed interactions and 10%, 20% and 30% of content with highest visibility on the platform?
4. What are the "top 100" users with the highest average engagement rate? What is their geographic origin and demography? Do they include multiple account users? How does the platform deal with detecting multiple account users (as a violation of the real name policy)?
5. How many of these "superusers" have been reported or flagged for abusive behaviour/violating platform policies?
6. What proportion of content from these 'superusers' is reported, flagged or otherwise classified by the platform as violating platform policies?
7. What signals does the algorithm pick up to amplify or reduce content visibility (i.e. flagging certain content as "to be recommended" or "not to be recommended")?

HYPOTHESIS 3

Editorial choices boost, protect or suppress some users over others, which can lead to censorship of certain voices

Description

Recommender systems contribute to imbalances in civic discourse. Research shows that certain political, social, ethnic or national groups receive "preferential treatment" and benefit from algorithmic amplification more than others. At the same time other groups face "algorithmic barriers" when trying to influence public debate. This mechanism has a negative impact on the pluralism of public debate and exacerbates discrimination.

By consistently ranking certain content or its producers higher, platforms give visibility boosts to certain media outlets or voices leaning towards one side of a political scene or conflict. Meanwhile, platforms suppress visibility of content which, for various reasons, they deem less desirable.

Users affected by measures suppressing visibility of their content have no way of effectively questioning them. These "algorithmic decisions" are taken in a non-transparent and arbitrary manner. Researchers speculate that observed effects result from platforms' "editorial preferences" (built into their algorithms) or arbitrary content moderation practices.

Supporting Evidence

Investigation of TikTok's feeds conducted by Tracking Exposed measured the relative visibility of Israeli and Palestinian activist content on the ForYou Page (the main recommendation feed of the platform), with the aim of probing the mechanisms by which a shadowban is applied in different geopolitical contexts. The experiment found a reduced visibility of Palestinian activist content (regarding pro-Palestinian movement and activists accounts that openly discuss Israel's violence against Palestinian people and the region) on TikTok's ForYou Page and at the same time the promotion of Israeli Defence Force-related TikToks showing up from an Israeli VPN. The results of the research imply that the Israeli ForYou Page promotes local content and shadowbans Palestinian activists' accounts (Romano & Faddoul, 2022).

According to a crowdsourced academic audit of recommendations made by Google, Google News, Facebook, YouTube, and Twitter, a small number of content producers dominate the gateways to news and information about key local and national issues. The findings also show each platform having its own distinct editorial preferences, which results in the adoption of algorithms that prioritise certain types of content over others, with professionally produced news dominant on some platforms but not others, and politically conservative mainstays like Fox News being particularly recurrent (Nechushtai et al., 2023).

“A small number of content producers dominate the gateways to news and information about key local and national issues.

Internal documents obtained by netzpolitik.org and The Intercept suggest TikTok was suppressing access to videos created by disabled, overweight, impoverished or LGBT+ users. For example, moderators were instructed to mark people with disabilities as "Risk 4". This means that a video is only visible in the country where it was uploaded. Overweight, and LGBT+ users ended up on a list of "special users" whose videos were regarded as a bullying risk by default and capped in their reach – regardless of the content. Their videos were automatically capped with the "Auto

R" mark, so that they did not exceed a certain number of views (after meeting the imposed "limit", they automatically ended up in the "not recommended" category). TikTok would also limit the reach of videos showing e.g. "dilapidatedhousing", "slums", "cracked walls" and "disreputable decorations", "abnormal body shape", "ugly facial looks", dwarfism, "obvious beer belly", "too many wrinkles", "eye disorders", and many other "low quality" traits. Clips including these features were considered too "unattractive" to be recommended in the "For You" section of the app (Biddle et al., 2020; Köver & Reuter, 2019).

Questions for Online Platforms

1. How does the platform monitor certain user groups that are considered vulnerable or represent minorities? How does the platform engage with certain user groups to understand their concerns?
2. Does the platform perform specific actions on users belonging to these groups (e.g. labelling)? If so, are these labels taken into account by the recommender system?
3. According to internal platform's research, are there any groups (users representing certain views) whose visibility is consistently amplified or reduced by the recommender system? If so, how is this effect explained?
4. Does the platform maintain aggregate demographic statistics, which may show systemic amplification of one group versus the other?
5. If there are such groups, what are the signals (such as for example a high "ratio" or flagging by the moderator) that the algorithm is picking up on to amplify/ reduce their visibility?
6. (How) Does the platform inform users about imposing content moderation measures which reduce visibility of their content ("shadowbans")? Does it allow users to effectively* contest those measures (*meeting the criteria set in Article 17 and 20 of the DSA)?
7. Which features qualify certain content or actors as "non-recommendable" and therefore negatively affect their visibility? Are these features described in terms and conditions (as required by Article 14 of the DSA)?
8. Are content moderation measures, which lead to reducing visibility of certain content, documented in transparency reports (as required in Article 24 and 42 of the DSA)?
9. Does the platform introduce any countermeasures to balance the visibility of groups representing different views in public debate and what are these measures?

HYPOTHESIS 4

Exploiting people's data to personalise content in a way that harms their health and wellbeing, especially for minors and vulnerable adults

Description

In order to customise the online experience for their users, social media recommender systems rely on behavioural patterns. These patterns may reveal individual vulnerabilities such as addictions, eating disorders, body complexes, anxiety or depressive disorders. As a result, recommender systems end up exploiting individual vulnerabilities to maximise user engagement. They also create feedback loops that drive users into narrower selections of content, corresponding

to their vulnerabilities. Such content may not be dangerous per se: and may be entirely acceptable when considered in isolation, but becomes harmful if consumed by vulnerable individuals.

Falling into a "doomscrolling" trap (e.g. excessive exposure to self-harm, diet-related content or idealised body images) triggers "unhealthy" engagement, which negatively impacts users' wellbeing and may exacerbate their pre-existing mental health issues. There are no "brakes" that would prevent a vulnerable user from getting stuck in a pattern of scrolling for negative information (Stray et al., 2022).

Supporting Evidence

In October of 2022, following an investigation and inquest into the death of 14-year-old Molly Russell, H.M. Coroner Mr Andrew Walker found that Russell "died from an act of self-harm whilst suffering from depression and the negative effects of online content". The coroner found that the recommendation engines of Instagram and Pinterest ultimately led Russell to her death. The coroner concluded that Russell had access to adult content that should not have been available for a 14-year-old child to see. In the prevention of future deaths report, he said: "the way that the platforms operated meant that Molly had access to images, video clips and text concerning or concerned with self-harm and suicide, or that were otherwise negative or depressing in nature. The platform operated in such a way, using algorithms, as to result, in some circumstances, in binge periods of images, video clips and text, some of which were selected and provided without Molly requesting them. These binge periods, if involving this content, are likely to have had a negative effect on Molly. Some of this content romanticised acts of self- harm by young people." (Walker, 2022)

The negative effect on self-esteem was correlated with the average daily time of use of the platform.

A survey conducted by an Italian centre for eating disorders in childhood and adolescence among 78 patients investigated the use of TikTok among young people with eating disorders. It found that the TikTok algorithm frequently showed users content relating to eating disorders without them having to even search for it (such content was both actively searched by patients and also proposed by the platform in a significant number of cases – over 60 % of respondents were showed pro-eating disorder content and over 50 % were shown promoted anorexia-related behaviour content). 59% of the investigated patients said that they felt more insecure after viewing contents related to diet. The negative effect on self-esteem was correlated with the average daily time of use of the platform (Pruccoli et al., 2022).

The analysis conducted by the Center for Countering Digital Hate (CCDH) found that TikTok-recommended videos about mental health or body image were served to "standard teen accounts" every 39 seconds. Recommended content included dangerously restrictive diets, pro-self-harm content and content romanticising suicide to users who show a preference for the material, even if they are registered as under-18s (Center for Countering Digital Hate, 2022).

A Wall Street Journal Investigation found that TikTok's algorithm's expose users to harmful content including videos about self-harm, extremely harmful dieting and suicide. The Wall Street Journal's investigative team set up more than 100 automated TikTok accounts or bots, assigning them a certain age, location and a short set of interests. At first the bots were shown a range of content, but then the algorithm started zeroing in on what would keep them watching. The bots that lingered over weight loss and exercise videos were quickly served more, until these topics made up more than half of the bots' feed (WSJ, 2021).

Questions for Online Platforms

1. What input features (e.g, demography, behavioural signals) are taken into account by the recommender system and what "weights" are attributed to them?
2. How does the platform differentiate (categorise) behavioural signals that show positive and negative engagement (feedback loops)?
3. It is known that the Facebook advertising system uses labels to segment users (Hitlin et al., 2019). Does the platform assign labels that may reveal race or sexual orientation (for e.g, White Christian, Asian and Queer) or mental health (e.g depressed, anxious) to serve sponsored content?
4. What is the impact of explicit user feedback? In particular, how does the alteration of settings (such as "show me less/show me more") affect the mix of suggested content?
5. How often is the "rabbit hole" effect (users driven to narrower selections of content) observed on the platform and how is it measured?
6. What steps does the platform take to diversify recommendations? How is "more diverse" content defined and selected? How successful are these measures at disrupting the "rabbit hole" effect?
7. What conclusions come from internal research on mental well being of platform users? What behavioural signals are collected to detect risks to mental health?
8. If the platform conducted A/B tests that involve mitigating risk of promoting content harmful for mental health, what alternative input features, performance metrics and designs were used?
9. Which demographics are more vulnerable to binge-watch harmful content? Are certain (marginalised/vulnerable) groups disproportionately affected by (exposure to) harmful content?

HYPOTHESIS 5

Building in features that are designed to be addictive at the expense of people's health and wellbeing, especially for minors

Description

Positive social stimuli results in the release of dopamine, reinforcing whatever behaviour preceded it. Cognitive neuroscientists have shown that rewarding social stimuli – laughing faces, positive recognition by our peers, messages from loved ones – activates the same dopaminergic reward pathways (and takes advantage of our desire for social validation). Social media apps are designed to encourage frequent and/or extended engagement by triggering a variable reward schedule and optimising the balance of negative and positive feedback signals, until their use becomes habitual. Compulsive social media behaviour is reinforced by technological affordances that allow users to enjoy a frictionless browsing experience (such as endless newsfeeds, auto refill, and autoplay functions). The resulting addictive use patterns pose significant risks, especially to minors going through important developmental phases.

> Multiple studies found that recommender-based products may encourage addictive tendencies.

Supporting Evidence

Multiple studies found that recommender-based products may encourage addictive tendencies. For example, researchers found that validation from likes on social media stimulates the brain in a similar way to cocaine, although not as intensely. Every notification, whether it's a text message, a "like" on Instagram, or a Facebook notification, has the potential to offer a positive social stimulus and dopamine influx. Moreover, to implement an even more effective variable-ratio reward schedule, Instagram introduced a notification algorithm which sometimes withholds "likes" on users' content, in order to deliver them in larger bursts. When a user creates a post, they initially may be disappointed to find fewer responses than expected, only to receive them in a larger batch later on. Their dopamine centres have been primed by those initial negative outcomes to respond robustly to the sudden influx of social appraisal (Haynes, 2018). Another study found that abstaining from social media use for a time, or allowing people to set future screen-time limits, produced a decrease in subsequent use, suggesting that social media use may indeed result in habit formation and self-control issues (Allcott et al.,2021).

Several studies concluded that excessive and problematic social media use – such as compulsive or uncontrollable use – has been linked to sleeping and attention problems. For example, a long-term study of several hundred Dutch teenagers showed that problematic social media use was significantly linked to the emergence of serious cognitive effects a year later, including reduced attention, increased impulsivity, and increased hyperactivity. Losing control over social media habits (such as lying to parents to gain access to social media) was significantly more likely to lead to new attentional problems a year later (Boer et al., 2020). A systematic review of 42 studies on the effects of excessive social media use also found a consistent relationship between social media use and poor sleep quality, reduced sleep duration, sleep difficulties, and depression among youth. Poor sleep has been linked to altered neurological development in adolescent brains, depressive symptoms, and suicidal thoughts and behaviours (U.S. Department of Health and Human Services, 2023).

Additionally, a number studies have shown that higher frequency of social media use has been associated with increases in depression, anxiety or neuroticism. Specific social media behaviours which increase odds of mental health problems include negative social comparisons (individuals who were more likely to compare themselves to others who are better off than them were also more likely to meet criteria for major depressive disorder) (Robinson et. al, 2019). It also includes social media-induced fear of missing out (concerns that others might be having rewarding experiences that one is absent from), which positively correlated with depression and anxiety symptoms (Fioravanti et al., 2021).

Questions for Online Platforms

1. What conclusions come from internal research regarding the scale and types of addictive behaviour on the platform? What categories of users (which demographics) are most vulnerable to develop addictive behaviour?
2. Does the platform analyse the impact of specific design elements on addictive use of the service?
3. What role do default notifications play in driving user engagement? How does the user engagement pattern change when notifications are disabled? How many users try this option?
4. How many users set screen-time limits for themselves? In how many cases is this additional safeguard circumvented by the user?

5. How many accounts belonging to minors are partnered with parental accounts?
6. How often are the parental control mechanisms used? In particular, how often do parents set screen-time limits for their children?

HYPOTHESIS 6

Using people's data to personalise content in ways that lead to discrimination

Description

Social media platforms have a commercial incentive to target ads to users in an increasingly narrow and direct fashion. Such personalisation requires collecting more and more data about users, including behavioural observations that may reveal (directly or "by proxy") sensitive characteristics. Ad delivery algorithms determine which users will see which ads. This process is distinct from ad targeting, which is determined by advertisers. Research shows that ad delivery algorithms create skewed (potentially discriminatory) outcomes in ways that advertisers do not intend. As a result, some users – due to their demographic, gender or racial characteristics – are less likely to see certain sponsored content (e.g. housing, employment or credit opportunities).This effect results from market and financial optimisation, as well as platform's own predictions about the "relevance" of ads to different segments of users.

Supporting Evidence

In a study examining the delivery of Facebook's employment and housing ads, researchers found that both the advertiser's budget and the content of the ad significantly contributed to the skew of Facebook's ad delivery along gender and racial lines. The daily budget of an ad impacted gender distribution of the audience, with higher budgets leading in general to a higher proportion of women being served the ads. In the context of the employment ads, despite the same bidding strategy, the same target audience and being run at the same time, researchers observed a significant skew along racial and gender lines due to the content of the ad alone. In the most extreme cases, ads for jobs in the lumber industry reached an audience that is 72% white and 90% male, ads for cashier positions in supermarkets reached an 85% female audience, ads for janitors were delivered to more than 65% women and 75% black users in aggregate, and ads for positions in taxi companies reached a 75% black audience. Similarly, in the context of housing ads, researchers found a significant ad delivery skew along racial lines, with certain ads being delivered to an audience of over 72% black users while others were delivered to an audience of as little as 51% black users (in principle houses for sale were being shown more often to white users and houses for rent were being shown more often to black users). Their research demonstrated mechanisms that can lead to potentially discriminatory ad delivery, even when advertisers set their targeting parameters to be highly inclusive (Ali et al., 2019).

“Both the advertiser's budget and the content of the ad significantly contributed to the skew of Facebook's ad delivery along gender and racial lines.

Another study, focusing on the distribution of "problematic" advertising among different age, gender, racial and ethnic groups, found that the ad diets of older users, born before 1980, are composed of 5.1% more of these ads than younger participants (with particular prevalence of "clickbait" and "scam" advertising which is shown more often to the older group). In this case, those differences existed both due to advertisers' targeting and the platform's ad delivery process – which, together, may create a feedback loop. Researchers also identified instances where the overall outcomes were different than delivery optimisation biases: black participants saw

a higher fraction of "clickbait" ads, but only when targeted by advertisers. On the other hand, Hispanic participants had higher exposure to "deceptive" ads, but only within ads that are essentially untargetted by advertisers, suggesting this effect was due to the platform's ad delivery process. Further, the research found that financial ads were shown more often to participants who identify as men, both as a system-level outcome, and when controlling for ad targeting. This included exposure to problematic financial products, but also financial opportunities (Ali et al., 2023).

Questions for Online Platforms

1. How coupled are the algorithms responsible for ranking organic content and recommending sponsored content? What percentage of all recommended content is sponsored content?
2. What are the input features of the algorithms recommending sponsored content? How is their performance measured (i.e. what is the definition of success and/or optimisation goal)?
3. How are content and users categorised (labelled) and clustered for the purpose of ad delivery? How does the platform ensure these categorisations are not based on sensitive personal data? How does the platform ensure that certain categories are not quasi-identifiers (proxies) for sensitive characteristics?
4. What measures have been implemented by the platform to prevent harmful and discriminatory categorisations based on behavioural patterns (inferred/observed data)?

SECTION 2

Key information for RAs and independent auditing

In addition to carrying out their own risk assessments with due diligence, VLOPs and VLOSEs should be obliged to disclose information that allows independent researchers to assess how platforms are mitigating against observed harms, and to better understand root causes of these harms:

1. High-level architectural description of the algorithmic stack

Modern recommender systems are typically not comprised of a single algorithm, but rather a series of algorithmic components, each with a specific function (e.g. selecting a pool of potential candidates to be displayed in the feed, scoring their likelihood to generate user engagement, diversifying the feed, filtering out potentially harmful content).

Platforms should disclose the overall architecture of their recommendation systems, for instance with a flow chart illustrating the different algorithmic modules, and their function in the recommendation process. For each of these modules it should then be possible to ask for further specifications, and for the role they might play in reinforcing or mitigating some of the risks to be outlined in this architecture document.

2. Technical specification of each algorithmic module:

a. What is the type of algorithm and its hyperparameters? For instance, for a neural network, what is the size and number of layers?

b. What are the various input features? For instance: user interaction history, content-specific features, author or channel-specific features, contextual features (device information, browser data, time of the day etc.). For each of these features, what is their relative importance? For instance, what are their weights in a linear regression model? For deep learning algorithms, platforms can use SHAPley or an interpretability model.

c. How are these features represented? If embedding spaces are used, what is the logic behind their training?

d. What is the loss function of the model? What are its different components, in a human interpretable way? What weights are given to these components (for instance balancing user and content creator value)?

e. How is the performance of the model measured at run time?

f. In the case of supervised and semi-supervised models, what constitutes the training data? What is the order of magnitude of data points? How often is the system re-trained or updated?

g. Is training data human-labelled? If so, are external partners (e.g. fact checkers) involved in the labelling process?

3. **How interpretable are the decisions made by the recommendation algorithm, and what efforts are made for interpretable machine learning pipelines?** What in-house algorithmic interpretability tools are used? What are the results of using these tools with respect to the feature importance and thresholds for various measures against violence, disinformation and hate speech? How accurate are the explanations given?

4. **When did the platform use specific measures to mitigate algorithmic harms** (such as "Break the Glass" measures known to be used by Meta in critical moments, to prevent spreading hate or disinformation) and what results came out of each intervention?

The information listed above should be disclosed for all classes of algorithms used in the recommender system, in particular: algorithms ranking organic content, algorithms ranking sponsored content and algorithms responsible for borderline content mitigation. This level of transparency would allow independent researchers and auditors to verify hypotheses formulated in this brief and many other papers exploring the relationships between harms experienced by social media users and key features used in recommender systems. For example, we could assess the efficacy of content moderation algorithms as well as the extent to which borderline content is allowed or recommended on the platform and the extent to which user behaviour influences recommendations.

Revealing key choices made by VLOPs and VLOSEs when designing their recommender systems would provide a "technical bedrock" for better design choices and policy decisions aimed at safeguarding the rights of European citizens online.

Borderline content: Content that is not prohibited on the platform, but comes close to the demarcation line.

Classifier/Detection Algorithm: A machine learning model that sorts input data into categories. Classification algorithms are supervised learning methods that predict a categorical response. Detection algorithms, similarly, predict whether a certain condition or feature is present (detected) or not. Examples include decision trees, support vector machines, and neural networks.

Collaborative Filtering: A technique used in recommendation systems that predicts the interests of a specific user by collecting preferences or taste information from many users. The assumption of this approach is that if user A has the same opinion as user B on a set of items, A is more likely to have B's opinion for a given item than that of a randomly chosen user.

Deep Learning: A subset of machine learning that is based on artificial neural networks with multiple layers (hence the "deep" in the name). These models are capable of learning from data that is unstructured or unlabeled, and they're exceptionally good at identifying patterns or features from input data. In the content of recommendation systems these might be used to choose features that are not explicitly fed into the algorithm, and to increase the "performance" of the feed.

Explicit/Implicit User Feedback: This refers to direct input from a user about their preferences. Examples include rating a movie on a scale of 1 to 5, liking or disliking a post, writing a review for a product, or answering a survey. This feedback is clear and intentional. In contrast, implicit user feedback comes from behavioural signals such as:

1. engagement with likes and shares,

2. click-through rate (the ratio of users who click on a recommended item to the total number of users who view the recommendation),

3. dwell time (the amount of time a user spends on a recommended item, such as watching a video or reading an article),

4. conversion rate (the ratio of transactions to the total number of sessions),

5. retention rate (how consistently users return to the platform),

6. session length (the total time a user spends on the platform during each visit).

Input Features: Also known as predictors or independent variables, these are the variables in a dataset that are used as input for machine learning models. They represent the characteristics or attributes of the data that the model will use to learn and make predictions. In the context of recommendation systems, this could include user behaviour data, reactions to posts (comments, sharing), time spent on an image or video, inferred user demographic data or label, user interactions in their network, or the text generated by the user.

Optimisation Goals: These refer to the various goals that most recommendation system algorithms optimise for, and can include the clickthrough rate (ratio of users who click on recommended item to the total who view the item), the dwell time, conversion rate (for advertisements), retention rate (how often users come back), session length and like/share/comment counts.

The next three terms refer to the technical makeup that allows the algorithm to reach its optimisation goals.

Loss Function: A method of evaluating how well a specific algorithm models the given data. It quantifies the disparity between the predicted and actual outcomes in the form of a single real number. Minimising the loss function is the main objective during the training phase of a model.

Optimizer Function: An algorithm or method used to adjust the parameters of a machine learning model to minimise the error (loss) of the model's predictions. Examples include Stochastic Gradient Descent (SGD), Adam, and RMSProp.

Performance / Accuracy: In the context of machine learning, performance or accuracy is a measure of a model's predictions against the true values for a given dataset. It is typically expressed as a percentage, where a higher percentage indicates a better fit between the model's predictions and the true values.

Organic vs algorithmic consumption: Organic consumption refers to the consumption of content users specifically asked for (e.g. they searched for or subscribed to), while algorithmic consumption refers to the consumption that is driven by recommendations (e.g. "Watch Next").

Rabbit Hole Effect: The rabbit hole effect refers to the phenomenon where users are continually served content that leads them deeper into a specific topic or viewpoint, often becoming more extreme or polarised in the process. This is often a result of recommendation algorithms that promote increasingly engaging (and often more extreme) content to keep users on the platform.

Ali, M., Goetzen, A., Mislove, A., Redmiles, E. M., & Sapiezynski, P. (2023). *Problematic Advertising and its Disparate Exposure on Facebook* (arXiv:2306.06052). arXiv. https://doi.org/10.48550/arXiv.2306.06052

Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2019). *Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes. Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW), 1–30. https://doi.org/10.1145/3359301

Allcott, H., Braghieri, L., Eichmeyer, S., & Gentzkow, M. (2020). The Welfare Effects of Social Media. *American Economic Review*, 110(3), 629–676. https://doi.org/10.1257/aer.20190658

Amnesty International. (2022). *The social atrocity: Meta and the right to remedy for the Rohingya*. https://www.amnesty.org/en/documents/asa16/5933/2022/en/

Biddle, S., Ribeiro, P. V., & Dias, T. (2020, March 16). Invisible Censorship: TikTok Told Moderators to Suppress Posts by "Ugly" People and the Poor to Attract New Users. *The Intercept*. https://theintercept.com/2020/03/16/tiktok-app-moderators-users-discrimination/

Boer, M., Stevens, G., Finkenauer, C., & van den Eijnden, R. (2020). Attention Deficit Hyperactivity Disorder-Symptoms, Social Media Use Intensity, and Social Media Use Problems in Adolescents: Investigating Directionality. *Child Development*, 91(4), e853–e865. https://doi.org/10.1111/cdev.13334

Brown, M. A., Nagler, J., & Tucker, J. (2021, October 27). Twitter amplifies conservative politicians. Is it because users mock them? *Washington Post*. https://www.washingtonpost.com/outlook/2021/10/27/twitter-amplifies-conservative-politicians/

Center for Countering Digital Hate. (2022). *Deadly by Design: TikTok pushes harmful content promoting eating disorders and self-harm into young users' feeds*. https://counterhate.com/wp-content/uploads/2022/12/CCDH-Deadly-by-Design_120922.pdf

Cunningham, T. (2023, May 8). *Ranking by Engagement*. https://tecunningham.github.io/posts/2023-04-28-ranking-by-engagement.html

Fioravanti, G., Casale, S., Benucci, S. B., Prostamo, A., Falone, A., Ricca, V., & Rotella, F. (2021). Fear of missing out and social networking sites use and abuse: A meta-analysis. *Computers in Human Behavior*, 122, 106839. https://doi.org/10.1016/j.chb.2021.106839

HateAid. (2021). *Boundless hate on the internet – Dramatic situation across Europe*. https://hateaid.org/wp-content/uploads/2022/04/HateAid-Report-2021_EN.pdf

Haynes, T. (2018, May 1). Dopamine, Smartphones & You: A battle for your time. *Science in the News*. https://sitn.hms.harvard.edu/flash/2018/dopamine-smartphones-battle-time/

Hindman, M., Lubin, N., & Davis, T. (2022, February 10). Facebook Has a Superuser-Supremacy Problem. *The Atlantic*. https://www.theatlantic.com/technology/archive/2022/02/facebook-hate-speech-misinformation-superusers/621617/

Hitlin, P., Rainie, L., & Olmstead, K. (2019, January 16). Facebook Algorithms and Personal Data. *Pew Research Center: Internet, Science & Tech*. https://www.pewresearch.org/internet/2019/01/16/facebook-algorithms-and-personal-data/

Huszár, F., O'Brien, C., Belli, L., Schlaikjer, A., & Hardt, M. (2021). *Algorithmic Amplification of Politics on Twitter*. Twitter. https://cdn.cms-twdigitalassets.com/content/dam/blog-twitter/official/en_us/company/2021/rml/Algorithmic-Amplification-of-Politics-on-Twitter.pdf

Köver, C., & Reuter, M. (2019, December 2). Discrimination: TikTok curbed reach for people with disabilities. *netzpolitik.org*. https://netzpolitik.org/2019/discrimination-tiktok-curbed-reach-for-people-with-disabilities/

Merrill, J. B., & Oremus, W. (2021, October 26). Five points for anger, one for a 'like': How Facebook's formula fostered rage and misinformation. *Washington Post*. https://www.washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/

META. (n.d.). About quality ranking. *Meta Business Help Centre*. https://en-gb.facebook.com/business/help/303639570334185

Mozilla. (2021). *YouTube Regrets. A crowdsourced investigation into YouTube's recommendation algorithm*. https://foundation.mozilla.org/en/youtube/findings/

Nechushtai, E., Zamith, R., & Lewis, S. C. (2023). More of the Same? Homogenization in News Recommendations When Users Search on Google, YouTube, Facebook, and Twitter. *Mass Communication and Society*, 0(0), 1–27. https://doi.org/10.1080/15205436.2023.2173609

Pruccoli, J., De Rosa, M., Chiasso, L., Perrone, A., & Parmeggiani, A. (2022). The use of TikTok among children and adolescents with Eating Disorders: Experience in a third-level public Italian center during the SARS-CoV-2 pandemic. *Italian Journal of Pediatrics*, 48(1), 138. https://doi.org/10.1186/s13052-022-01308-4

Reset & Hate Aid. (2022). *9 months after the Capitol Hill insurrection, Big Tech puts the German election at risk*. https://hateaid.org/wp-content/uploads/2022/04/210831_Reset_Facebook_Bundestagswahl_EN.pdf

Ricks, B., & McCrosky, J. (2022). *Does This Button Work? Investigating YouTube's ineffective user controls*. Mozilla. https://foundation.mozilla.org/en/research/library/user-controls/report/

Robinson, A., Bonnette, A., Howard, K., Ceballos, N., Dailey, S., Lu, Y., & Grimes, T. (2019). Social comparisons, social media addiction, and social interaction: An examination of specific social media behaviors related to major depressive disorder in a millennial population. *Journal of Applied Biobehavioral Research*, 24(1), e12158. https://doi.org/10.1111/jabr.12158

Romano, S., & Faddoul, M. (2022). *Mapping Ban and Shadow-Ban on TikTok: Expose hidden censorship with a cross-national research*. https://tiktok.tracking.exposed/ws22-shadowban-research/

Seetharaman, D., & Horwitz, J. (2020, May 26). Facebook Executives Shut Down Efforts to Make the Site Less Divisive. *Wall Street Journal*. https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499

Stray, J., Halevy, A., Assar, P., Hadfield-Menell, D., Boutilier, C., Ashar, A., Beattie, L., Ekstrand, M., Leibowicz, C., Sehat, C. M., Johansen, S., Kerlin, L., Vickrey, D., Singh, S., Vrijenhoek, S., Zhang, A., Andrus, M., Helberger, N., Proutskova, P., … Vasan, N. (2022). *Building Human Values into Recommender Systems: An Interdisciplinary Synthesis* (arXiv:2207.10192). arXiv. https://doi.org/10.48550/arXiv.2207.10192

U.S. Department of Health and Human Services. (2023). *Social Media and Youth Mental Health: The U.S. Surgeon General's Advisory*. https://www.hhs.gov/surgeongeneral/priorities/youth-mental-health/social-media/index.html

Walker, A. (2022). *Regulation 28 Report To Prevent Future Deaths*. North London Coroner's Service. https://www.judiciary.uk/wp-content/uploads/2022/10/Molly-Russell-Prevention-of-future-deaths-report-2022-0315_Published.pdf

WSJ. (2021, July 21). Inside TikTok's Algorithm: A WSJ Video Investigation. *Wall Street Journal*. https://www.wsj.com/articles/tiktok-algorithm-video-investigation-11626877477

Zuckerberg, M. (2021, May 5). *A Blueprint for Content Governance and Enforcement*. Facebook. https://www.facebook.com/notes/751449002072082/